Univerzita Karlova v Praze
Filozofická fakulta

Diplomová práce

2010 Jiří Milička

Univerzita Karlova v Praze
Filozofická fakulta
Ústav Blízkého východu a Afriky

Diplomová práce

Jiří Milička

Kvantitativní pohled na strukturu arabského textu

Quantitative View on the Arabic Text Structure

2010

Vedoucí práce: Doc. Petr Zemánek

Děkuji svým rodičům za plnou podporu, které se mi od nich dostalo, a to nejen po materiální stránce.

Děkuji vedoucímu této práce, docentu Petru Zemánkovi, za odvahu, s níž přijal toto poněkud nezvyklé téma, a za cenné připomínky, které pomohly ke vzniku této studie.

Dalším pedagogům Filozofické fakulty Univerzity Karlovy.

A múze, která mě nezištně inspirovala.

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

V Praze dne 24. 8. 2010 Jiří Milička

Anotace

Tato studie navrhuje několik obecných kvantitativně lingvistických falzifikovatelných hyptotéz, které následně ověřuje na textech v moderní spisovné arabštině, klasické arabštině a několika evropských jazycích, přičemž arabština a čeština slouží také jako zdroj inspirace. Hypotézy se týkají struktur, které vytvářejí frekvence a délky slov ve větách a v nadvětných celcích, dále souvislostí vztahu délek vět a frekvencí slov v nich obsažených s Menzerathovým-Altmannovým zákonem a pohledu na text pomocí tzv. kombinatorického zobrazení.

Klíčová slova

arabština, arabský jazyk, lingvistika, kvantitativní lingvistika, type-token relation, délka slova, délka tokenu, frekvence slova, frekvence typu, struktura věty, nadvětná struktura

Abstract

The thesis suggests several general quantitative linguistic falsifiable hypotheses and tests them on corpora of standard modern Arabic, medieval Arabic and some European languages, including Czech and English. The hypotheses deal with structures built by word lengths and word frequencies within sentences and supra-sentential elements, with connection between sentence length – its constiuents frequency relation and Menzerath-Altmann Law, and with a view on text via so-called combinatorial mapping.

Key Words

Arabic, Arabic language, linguistics, quantitative linguistics, type-token relation, word length, token length, word frequency, type frequency, sentence structure, suprasentential structure

Obsah

1. Úvod

Zkoumání jazyka vyžaduje pokoru. V textu se odráží vnější i vnitřní svět mluvčího a jen velmi těžko můžeme rozlišit, zda jsou jeho určité vlastnosti dány předmětem, o kterém tento text pojednává, způsobem, jakým člověk o tomto předmětu přemýšlí, nebo způsobem, jakým se tyto myšlenky dostávají na povrch, aby jim byli schopni porozumět další lidé[1].

To je také důvod, proč v této studii spíše než se slovem „jazyk", budeme operovat se slovem „text"[2].

Naším cílem je formulování obecně platných zákonů. Tedy nalézání testovatelných hypotéz[3] týkajících se všech textů. Při hledání obecných zákonů je zcela zásadní intuice, neboť jakkoli systematický přístup nemůže nahradit onu poněkud tajemnou schopnost lidského mozku, která z omezených informací dokáže pomocí analogie a neúplné indukce vytvářet zcela nové hypotézy. A právě zde tkví zásadní postavení arabštiny v naší práci, neboť při znalosti pouze jednoho jazyka nás intuice bude svádět k hypotézám nedostatečně obecným, které se sesypou, když je aplikujeme na texty v různých světových jazycích. Znalost rodné češtiny a jiných indoevropských jazyků tvoří první pilíř, o který se bude opírat naše uvažování. Druhým pilířem bude arabština, jazyk, který nemá na indoevropskou rodinu zjevné genetické vazby a jehož areální interakce s češtinou byla, navzdory geografické blízkosti Evropě, minimální, a to zejména kvůli

---

[1] Myslím, že celá kontroverze kolem věty „myslíme jazykem", kterou ještě občas slýcháváme, je dána nejasností hranic, kde je myšlenka formulována do sdělitelné podoby, nebo spíše tím, že sami na sobě nejsme schopní tuto hranici pozorovat a do jiných mozků nevidíme. Můj subjektivní pocit je, že při přemýšlení o konkrétních pojmech se úpravou do sdělitelné podoby člověk nezdržuje, ale naopak při vyšší abstrakci si člověk formuluje text zcela samovolně. Ovšem orwellovská idea, že zakážeme-li užívání slova „svoboda", pak lidé přestanou o svobodě přemýšlet, je poněkud scestná – uvědomíme-li si, že přemýšlet můžeme i o věcech, které dosud jméno nemají. I v této studii najdeme několik pojmů, které byly před jejím napsáním zcela neznámé a které nemají ekvivalent v jiném jazyce než v češtině.

[2] Laikovi se to může zdát neuvěřitelné, ale pojem *jazyk* nemá v lingvistice pevnou definici. U nás se obvykle, podle strukturalistické tradice, hovoří o jazyce jakožto *langue* jako o souboru konvencí, pomocí kterých se řídí artikulovaná řeč (od které je odvozen její písemný záznam). Osobně (a snad v souladu se čtenářem) definuji jazyk raději jako metodu, pomocí které může sdělit člověk svou myšlenku jinému člověku (materialisticky řečeno „popsat stav svého mozku"). Hádky o definice považuji za neplodné, ovšem je možné, že se toto mé pojetí nějak zobrazí v následujícím textu a rád bych předešel možným nedorozuměním. Definice *textu* v tomto širokém pojetí pak zní: Text je výsledkem snahy člověka o sdělení svých myšlenek jinému člověku. Pro další definice viz kapitolu 1.6.

[3] V popperovském smyslu, což rozebírám v kapitole Epistemologická východiska.

náboženským a kulturním rozdílům. Podobně pomohla turečtina Luďku Hřebíčkovi k tomu, aby mohl rozvinout Altmann-Menzerathův vztah do komplexního systému. I když náš výzkum bude zcela nezávislý na Hřebíčkových textových teoriích, nemůžeme zapřít jistou inspiraci jeho populární knihou Vyprávění o lingvistických experimentech s textem[4]. Se zmínkou o této knize se dostáváme ke třetímu pilíři, o který se bude opírat naše intuice: data z různých měření, která jsme během naší práce provedli. Myšlenka kvantifikace vlastností textu je velmi stará, avšak dodnes není samozřejmá. Nutno dodat, že tato data sama o sobě sice velký pokrok ve vědě nepředstavují, ale postup vpřed by byl bez nich nepředstavitelný[5]. Samozřejmě je možné předkládat testovatelné hypotézy týkající se obecně textu i bez kvantifikace, ovšem v historii lingvistiky najdeme jen velmi málo, pokud vůbec, testovatelných hypotéz týkajících se textu nebo jazyka obecně, které by neměly kvantitativní povahu[6]. Pokud bychom uznali rozhodující úlohu kvantifikace ve schopnosti vědy přinášet testovatelné hypotézy (což je v přírodních vědách neoddiskutovatelné), pak je jasné, že obtížné měření vlastností textu a technické nároky na práci s informacemi zabraňovaly lingvistům před rozšířením počítačů ve skutečném testování jejich hypotéz a tak, možná zbytečně kriticky řečeno, stanovování axiomů, zkoumání singularit a diskuze o definicích udržovaly tuto vědu při životě.

---

[4] (Hřebíček, 2002); tato kniha mě ovlivňovala v období dospívání a formovala můj pohled na jazyk a lingvistiku jakožto vědu, která je teprve na začátku úchvatné cesty za poznáním základních zákonů, kteroužto etapu mají ostatní vědy již dávno za sebou. Ovšem nemůžeme se tomuto stádiu lingvistiky divit, ono okřídlené „tření zanedbejte“, které známe z učebnic fyziky, je ozvěnou tradiční metody experimentů, při které se fyzikové snažili a snaží oddělit různé mechanizmy, které působí na konečný výsledek (o to zajímavější je pro nás například Galileiho myšlenkový experiment, kterým odvodil princip volného pádu ve vakuu, aniž by měl k dispozici vývěvu). Lingvistika má velmi omezené možnosti tyto různé mechanizmy oddělovat a metoda, kterou si budeme moci vyzkoušet v následujících kapitolách, není použitelná za všech okolností.

[5] Je otázka, na kolik má smysl tato data, vlastně jednotlivá singulární pozorování, publikovat. Možná jen pokud bylo k jejich vypracování potřeba velké množství energie, jako například v případě některých publikací ÚČNK, neboť přebírání cizích dat pro vlastní potřebu je poměrně náročné (heuristika někdy zabere více času než měření dat osobně) a navíc vnáší do konečného výsledku nejistotu ohledně techniky měření a měřeného vzorku. Je tedy možné, že mnohá měření podobná těm, která jsem provedl, již publikoval někdo přede mnou, aniž bych o tom věděl a aniž bych ho citoval. Ovšem hypotézy, které předkládám, by měly být všechny originální.

[6] Ke své vlastní škodě jsem jen velmi málo zběhlý v dějinách lingvistiky, konzultoval jsem tedy tuto otázku s kolegy, například s Janem Chromým, ovšem ani pak se nepodařilo najít jediný příklad jednoznačně vyvratitelného obecného jazykového zákona, který by byl vyjádřen před tím, než vstoupil na scénu George Kinsley Zipf (1932).

V této práci nepoužijeme jedinou matematickou operaci nebo metodu měření, která by v Evropě nebyla známa již v předminulém století, a podle mého hrubého odhadu, kdyby se v roce 1859 francouzští investoři rozhodli, že by jim tato studie přinesla větší užitek než stavba Suezského průplavu, nic by nebránilo tomu, aby byla provedena všechna měření a výpočty, kterých bylo zapotřebí. Avšak teprve lehkost, s jakou můžeme díky počítačům pořádat výpravy do neprobádaných končin, ze kterých nečekáme žádný zisk, nám dává možnost v těchto končinách nalézt něco zajímavého.

1.1 Stav výzkumu

Čistě kvantitativní studie zkoumající arabštinu jsou vzácné – ještě před deseti lety Susan Abee (2000) tvrdila, že nikdo před ní nezměřil průměrnou distribuci arabských slov podle délky. Při heuristickém bloudění různými časopisy, které se specializují na semitistiku, a sborníky semitistických konferencí spíše nalezneme ad hoc použité metody komputační lingvisticky převzaté z evropských jazyků, často upravené tak, aby lépe fungovaly na semitských jazycích[7], totéž se týká časopisů specializovaných na komputační lingvistiku[8]. Hlavní důraz je kladen na praktické využití, jedná se spíše o pragmatická řešení různých úkolů při zpracování textu nebo budování korpusů, která využívají znalostí z filologie. Stav bádání v arabském světě dokládá fakt, že zatímco generativní lingvistika (الالسنية التوليدية) a transformační gramatika (الالسنية التحويلية) si své stoupence našly, něco takového jako pojem kvantitativní lingvistika (الالسنية المقدارية, nebo snad الالسنية الكمية) v arabském odborném diskurzu budeme hledat jen těžko. Celkový dojem z literatury na tomto poli je, že se komputační lingvisté často pokoušejí o vymýšlení algoritmů na zpracování jazyka, aniž by se snažili zjistit, jak vlastně funguje, bez snahy o hlubší pohled do struktury arabského textu na základě kvantitativního zkoumání[9].

Co se ostatních blízkovýchodních jazyků týká, lepší situace je v perštině, kterou již začali kvantitativně zkoumat sami Íránci, a také v turečtině, mimo jiné proto, že je tradičním hájemstvím Luďka Hřebíčka. Ten využívá turečtiny ke svým obecně lingvistickým kvantitativním experimentům, neboť je původním zaměřením turkolog. Využívá přitom výhodnou vlastnost tureckých slov – a totiž že je lze rozdělit na morfy téměř automaticky bez přispění člověka,

[7] Například morfologická analýza Otakara Smrže (Smrž 2007). Podobným tématem se zabývá George Kiraz (Kiraz 2001) a Ken Beesley (Beesley 1996).

[8] Například v časopise Computer Speech and Language (Alnajem, 2006), (Kirchhoff a kol., 2006).

[9] Podle vžité terminologie korpusové lingvistiky bychom tento přístup označili jako *corpus based approach* bez snahy o *corpus driven approach.*

ovšem myslím, že hlavním důvodem je, že perfektní znalost jazyka, který není příbuzný s rodným jazykem, dává lingvistovi schopnost nahlížet na text s větším nadhledem. Podobně Gabriel Altmann[10] je původně indonesista a japanolog. Také já se budu opírat o arabštinu hlavně z těchto důvodů, jak jsem již předeslal v úvodní kapitole.

Tvrdé jádro komunity kvantitativních lingvistů tvoří několik (často emeritních) profesorů a lidí seskupených kolem nich, působících spíše na provinčních univerzitách (Trier, Bochum, Trnava, Graz…). Kvantitativní lingvistika totiž nemá téma, které by přitahovalo granty (soužití menšin, globální oteplování, bezpečnostní politika) ani sponzory (rychle zpeněžitelné technické inovace). Díky tomu (na rozdíl například od ekonomie, klimatologie, sociologie, politologie a farmacie) není vystavena žádným vnějším tlakům, a tak si může dovolit čistotu metody a oproštění od ideologií a dogmat. Zároveň jsou ovšem její teorie naostro testovatelné (na rozdíl třeba od takové literární vědy, která si na ekonomické tlaky také nemůže stěžovat). Pokládám proto studium této vědní disciplíny za ideální mentální cvičení a věřím, že energie vložená do této práce nebyla ztracena, i když poznatky v ní obsažené budou zapomenuty.

1.2 Oblast zájmu

Je zvykem dopředu vytyčit cíle studie a také my se přidržíme této konvence, přestože skutečný výzkum probíhal poněkud méně uspořádaně, než jak bude popsáno v této kapitole.

Nejprve se při práci zastavíme u dosud opomíjené skutečnosti, že průměrná četnost slova na různých pozicích ve větě není homogenní, například na koncích vět se častěji vyskytují vzácná slova než jinde ve větě. Při testování této hypotézy narazíme na struktury, které prozkoumáme důkladněji. Podobné struktury pak v další kapitole nalezneme i pro délky slov a také se na ně podíváme blíže a vzájemně je srovnáme.

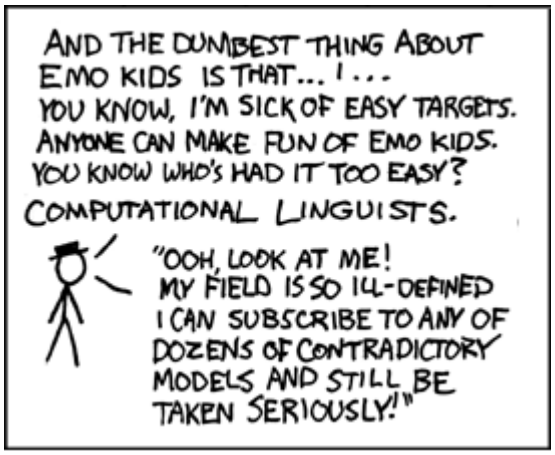
Necháme se inspirovat zjištěním, že průměrná frekvence slov ve větě závisí také na její délce a v následující kapitole se pokusíme nalézt souvislost mezi tímto vztahem a Menzerathovým-Altmannovým zákonem.

Dále pojmy frekvence a délky slov osvobodíme z hranic větných celků a v souladu se zadáním se budeme soustředit na kolokace a nadvětné struktury. Díky výsledkům tohoto bádání budeme moci zkonstruovat dvě zobrazení textu, pro která najdeme i praktické uplatnění. Poslední kapitolu pak zasvětíme kombinatorickému zobrazení, jež známe z předchozího bádání, srovnáme

[10] Kolem Gabriela Altmanna a nakladatelství RAM-Verlag se soustřeďuje největší aktivita kvantitativní lingvistiky, ať už různé časopisy na toto téma (Glottotheory, Glottometrics, Empirical Text and Culture Research), nebo encyklopedie Laws in Quantitatie Linguistics na webu university v Trieru (Altmann – Köhler – Vulanović 2006).

jeho výsledky se dvěma zobrazeními textu z předchozí kapitoly a podíváme se, jestli nám lépe neukazuje nadvětné struktury.

1.3 Epistemologická východiska



Uvedená karikatura[11] pěkně dokumentuje, co si o komputačních lingvistech myslí jejich kolegové na matematicko-fyzikálních fakultách ve světě. Myslím, že důvodem této špatné pověsti je metodologická a definiční neukotvenost, která by v jiných vědách byla stěží myslitelná[12], a nejasná hranice mezi empirickou vědou, normativními texty a planým verbalizmem. Karl Raimund Popper (1997: str. 18) stanovuje jasné demarkační kritérium mezi vědou a metafyzikou (což je dost eufemistické pojmenování, neboť do druhé zmíněné kategorie řadí všechny výroky, které nesplňují jeho podmínky vědeckosti), přičemž metafyziku nezavrhuje, pouze odděluje. Tím kritériem je falzifikovatelnost hypotéz: máme obecné tvrzení (například všechny vrány jsou černé) a toto tvrzení nemůžeme verifikovat (prozkoumat všechny vrány ve vesmíru), ale můžeme ho falzifikovat (najít alespoň jednu vránu, která není černá). *Singulární tvrzení*, které může hypotézu falzifikovat (existuje alespoň jedna vrána, která není černá) musí být *intersubjektivně* verifikovatelné – u intersubjektivního testování se zastavíme v následující podkapitole.

---

[11] A nejlegračnější věc na příznivcích EMO je... No, malé cíle mě nudí, z EMO si může dělat legraci každý. Víte, kdo to měl vždycky až moc lehké? Komputační lingvisti. „Podívej, moje oblast výzkumu je tak blbě definovaná, že se můžu přihlásit ke kterékoli ze všech vzájemně si protiřečících teorií a pořád mě budou brát vážně." Dostupné z: http://xkcd.com/114/ [1.7.2010].

[12] Představme si lékaře, který své hypotézy ověřuje pouze sám na sobě pomocí *introspekce*, nebo fyzika, který každých pár desítek let změní od základů názvosloví. Na druhou stranu musíme lingvistice přiznat těžko uchopitelný předmět bádání.

Tím se Popperovi podařilo vyhnat do „metafyziky" výroky jako „když strom padá, slabě sténá, ale jen když nikdo neposlouchá", nebo „všichni andělé mají křídla". Zároveň ovšem nepřiznává statut vědecké hypotézy ani výrokům typu „existuje prvek s protonovým číslem 72"[13]. Proto bychom Popperovi mohli vytknout přílišné přeceňování univerzálních hypotéz, tedy vlastně non-existenciálních tvrzení[14] na úkor tvrzení existenciálních (existuje černá vrána) – tím by se například velké části zmiňované geografie nebo historiografie dostaly za hranici empirické vědy[15]. Nemůžu vyloučit, že právě tam, vedle filozofie, teologie, právní „vědy" a magie, by Popper historii rád viděl, nicméně myslím, že se shodnu se čtenáři, že pro naše potřeby je vhodnější poněkud univerzálnější demarkační kritérium:

Za vědeckou hypotézu považujeme každé intersubjektivně vyvratitelné non-existenciální tvrzení a každé intersubjektivně ověřitelné existenciální tvrzení.[16]

Mezi empirické zákony můžeme také řadit zákony, které není možné testovat, ale lze je jako tautologii vyvodit z jiných testovatelných zákonů (tento status měla po dlouhou dobu Einsteinova teorie relativity).

1.3.1 Intersubjektivita testování

Popperova podmínka, aby každé testování mohlo být intersubjektivní (Popper 1997: str. 81-101), odstraňuje z vědy všechny výroky, které mohla testovat jen jediná autorita, které je nutno věřit. Přírodní vědy také na opakovatelných pozorováních nebo experimentech trvají nejméně od

---

[13] (Popper 1997: str. 54) Zde Popper doporučuje nějakým způsobem z existenciálních tvrzení vyvodit tvrzení obecná. Dovedu si představit, že například geografie, která je do značné míry na existenciálních tvrzeních postavená, namísto *singulárního tvrzení* „na 7°53'48" j. š. 140°33' 40" z. d. je nebezpečný útes" přednese *falzifikovatelnou hypotézu*, že „každá loď, která se pokusí proplout 7°53'48" j. š. 140°33' 40" z. d. ztroskotá". Účelnost takové práce nechám na posouzení čtenáři.

[14] „Každá vrána je černá" je ekvivalentní výroku „neexistuje vrána, která není černá".

[15] I když historiografie obsahuje obecné hypotézy (např. „žádný starý Egypťan neznal kolo" – toto tvrzení nemůžeme verifikovat, pouze falzifikovat, třeba tak, že pod Cheopsovou pyramidou najdeme starý žebříňák), jako historik mám pocit, že těžiště této vědy se nachází právě ve sbírání a ověřování oněch „singulárních tvrzení" (např. „staří Egypťané postavili Cheopsovu pyramidu").

[16] Tímto se nám podaří žabomyší války o normování spisovné češtiny, třeba střet tradiční *metody dobrého autora* a Cvrčkův *koncept minimální intervence* (Cvrček 2009), pochopit jako souboj ideologií a v rámci vědeckého diskurzu se jím nezabývat, aniž bychom tím zmenšovali jeho důležitost.

časů Francise Bacona. To je ovšem v některých případech těžko splnitelná podmínka (zřídka se opakující okolnosti pro pozorování, nákladné podmínky pro experiment) a v praxi jsme tak ve všech oblastech vědy nuceni věřit zákonům, které testoval někdo za nás. Zmiňovaná historiografie vyvinula důmyslnou metodiku ověřování existenciálních tvrzení, odhalování důvěryhodnosti autority, která o něm referuje, jakož i informačních kanálů mezi onou autoritou a námi[17].

1.3.2 Otázka definic

Snad jako atavismus zbyl lidstvu pocit, že kdo umí věci správně pojmenovat, poznal je a dokáže je pak lépe ovládat; asi vrcholem takových představ bylo, když středověcí alchymisté hledali nejmocnější Boží jméno, aby jeho pomocí ovládli přírodu. Myslím, že s tímto pocitem úzce souvisí myšlenka, že dokážeme-li nějaký pojem *správně* definovat, pak jsme pokročili v poznání předmětu, který tento pojem označuje. Víra, že vůbec existuje nějaká *správná* definice, kterou bychom se měli alespoň pokusit nalézt je zvláště v „humanitních vědách" hluboce zakořeněná[18]. Neodvažuji se tuto víru zcela popírat – jistě, pro každou hypotézu můžeme najít jinou adekvátní sadu definic (například tvrzení „každá vrána je černá" dovoluje, abychom pod pojem *vrána* počítali i některé mrtvé vrány, naopak tvrzení „vrány jsou všežravci" nepočítá s tím, že bychom do definice zahrnovali mrtvé vrány), ale z praktických důvodů je účelné, aby se našla a rozšířila pokud možno jedna zapamatovatelná definice, která by neomezovala hypotézy, ve kterých je užita. Na druhou stranu je třeba si uvědomit, že samotné stanovení definice bez vyslovení jakékoli hypotézy k našemu vědění nepřispívá.

K. R. Popper se zastavuje nad tím, že pokud bychom chtěli všechny pojmy definovat, tak se dostaneme do nekonečného regresu, definice podle něj mají místo tam, kde hrozí, že by se autor

---

[17] Je zajímavé, že sami přírodovědci si ve chvíli, kdy je narušena podmínka intersubjektivity, často nevědí rady, na rozdíl od „humanitně" vzdělaných lidí. Jako příklad uvedu Kremlíkův blog (Kremlík 2010) který je slabý, kdykoli se dotýká technické stránky globálního oteplování, avšak skvěle reflektuje, jakými metodami IPCC informuje veřejnost, práci médií s těmito informacemi, finanční a mocenské zájmy zainteresovaných osob a snahu, aby byla veškerá intersubjektivita testování hypotéz představených IPCC nemožná a byla nahrazena důvěrou v autoritu tohoto orgánu OSN. Nutno ovšem dodat, že podobnými neduhy „popíračů globálního oteplování" se autor blogu nezabývá. Těžko říci, jak dlouho bude trvat, než klimatologie přestane být válečným polem ideologií a vrátí se jí statut přírodní vědy.

[18] Tento fenomén se krásně ukazuje v 1. kapitole Kropáčkova *Islámského fundamentalismu* (Kropáček 1996), kde autor na několika desítkách stran srovnává různé definice pojmu „islámský fundamentalizmus".

se čtenářem neshodli na vymezení nějakého pojmu, přičemž předpokládá, že čtenář se s autorem *chce* shodnout.

V mnoha vědách mnoha pojmům rozumíme z kontextu, aniž bychom kdy slyšeli jejich definici prostě proto, že se jejich význam kryje s tím, jak jsou užívány v běžném jazyce. Během historie lingvistiky ovšem některé jinak běžné pojmy často získaly poněkud neintuitivní obsah, proto pro jistotu v kapitole 1.6 uvádím i definice pojmů, jako je *jazyk*, *text* a *slovo*.

1.3.3 Pravděpodobnostní zákony

Někdy je složité vysvětlit lidem, že když nějaké tvrzení platí pro průměrný prvek množiny, pak nemusí platit pro všechny prvky[19]. Zipf, který se s takovými nedorozuměními setkával zřejmě často, k tomu trefně poznamenává:

(…) statistics are hateful to the human mind; they are painfully definite for the group without being particularly definite for the individual. Undoubtedly, a primary law which knows no fluctuation within itself is pleasanter. If nature had consulted man in the latter, we should all have suggested primary laws. . . .but nature did not consult us . (Wyllys 1981: str. 4)

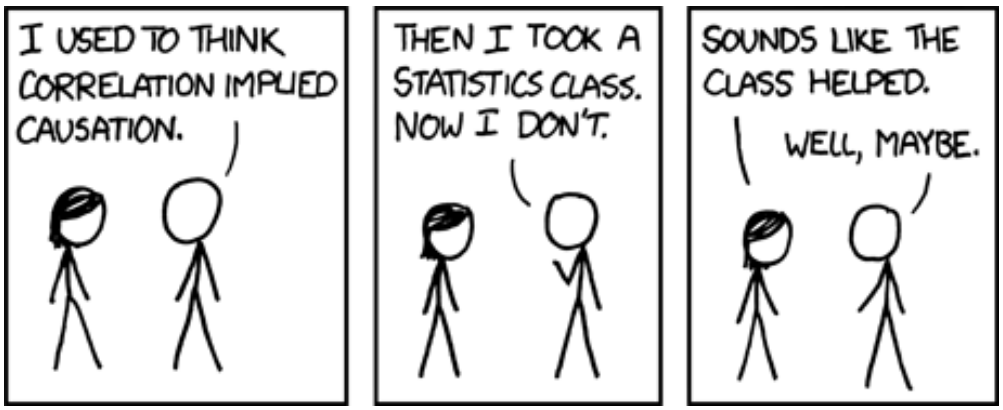
Popper uznává, že pravděpodobnostní zákony jsou těžko testovatelné (např. pravděpodobnost, že při hodu mincí padne panna je 1/2, ale není vyloučené, že při všech pokusech o testování této hypotézy nepadne ani jednou, neboť máme k dispozici konečný počet pokusů). Odmítá formulaci, že „zákon platí z určitého počtu procent" a raději uvažuje o tom, kudy vést hranici pro odmítnutí. (Popper 1997: str. 198). V praxi se ve „společenských vědách" za hranici významnosti pokládá 95 %, resp. 99,9 % (Volín 2007: str. 36), ta je však dána spíše tradicí a je lépe ji pragmaticky přizpůsobit účelu, za jakým hypotézu testujeme.

Při interpretaci pravděpodobnostních zákonů je třeba nezapomínat, že korelace nezakládá příčinný vztah[20]. Informace, že jev A koreluje s jevem B nám sama o sobě nesděluje nic o tom, jestli jev A ovlivňuje jev B, nebo obráceně, nebo jestli neexistuje nějaký jev C, který ovlivňuje jevy

[19] Kdysi jsem se bavil s jednou tradičně založenou filoložkou a když chtěla, abych jí dal příklad nějakého zákona kvantitativní lingvistiky, uvedl jsem hypotézu, že delší slova mají průměrně menší frekvenci než slova kratší. Velmi mě překvapila, když kontrovala, že to přeci není pravda a uvedla *protipříklad:* vždyť slovo „nebo" je delší než slovo „val", ale určitě je četnější…

[20] Také tato základní logická poučka byla v klimatologii opuštěna, často se zanedbává i v ekonomii a zejména v sociologii.

A i B. Vypadá to jako samozřejmost, ale média nás denně přesvědčují o opaku. K tomuto tématu mohu uvést karikaturu od stejného autora, jako na začátku kapitoly[21]:



Podobně se vyjadřuje také jedno arabské přísloví:

اذا صاح الزاغ والمطر نزل لا يعني أنّ الزاغ سبب المطر.[22]

1.4 Metodika

Metodika vychází z obecných principů, jak jsme je popsali v předchozích kapitolách, aplikovaných na každý problém zvlášť. Práce probíhala méně systematicky, než jsem si zprvu představoval.

Nejprve jsem vždy určil jazykovou strukturu, která z nějakého důvodu vypadala perspektivně (většinou na ni směřovaly otázky, které vyplynuly z předchozího výzkumu) a stanovil metodu, která by zjednodušila pohled na ni. Obvykle stačilo pouze kvantifikovat a změřit nějakou vlastnost textu. Ovšem často bylo nutné výsledky měření porovnat s modelem, který aproximoval, jak by měření dopadlo, kdyby proběhlo na „textu" zbaveném hledané struktury. Tyto matematické modely jsou v každé kapitole popsány.

Poté jsem vybral vhodný text, na který by bylo možno metodu aplikovat. Často bylo nutné měření opakovat několikrát, než z dat vyplynuly obecné zákonitosti.

Tyto obecné zákonitosti jsem se pak snažil co nejexaktněji formulovat a v rámci možností ověřit (říkám spíše „ukázat jejich platnost na názorném příkladu" než „dokázat jejich platnost") – tedy stanovit metodu falzifikace a aplikovat ji. Nejde o koroboraci teorie v pravém slova smyslu,

---

[21] Vždycky jsem si myslel, že korelace implikuje kauzalitu... Pak jsem absolvoval přednášky statistiky a už si to nemyslím... Takže přednášky pomohly... No, možná. Dostupné z http://xkcd.com/552/ [1. 7. 2010].

[22] Když zakrákorá vrána a spustí se déšť, neznamená to, že jej způsobila ona vrána. (Bahbouh 2001, str. 134)

avšak věřím, že pokud někdo bude stát o precizní zjištění, zda se dá na výstupy té které teorie spolehnout (například kdyby našel pro některou z nich využití, nebo na ní chtěl stavět další úvahy), bude schopný testy opakovat a rozšířit o další data. I v tom potenciálnímu zájemci samozřejmě nabízím pomocnou ruku.

1.5 Technika

Pro řešení konkrétních úkolů jsem využil východiska korpusové lingvistiky a techniky, které se mi osvědčily v minulosti. První zkušenosti s kvantitativním zpracováním textů jsem získával přibližně od roku 2002 při středoškolské odborné činnosti na téma *Využití internetu při určování eufoničnosti jazyků pomocí statistického rozboru textů*. Z této doby také pocházejí mé první počítačové programy na zpracování textu, jejichž některé části v různých podobách využívám dodnes. Byly tehdy napsány v didaktickém programovacím jazyce Pascal, a abych zachoval kontinuitu, využívám dnes programovací jazyk *Delphi*, který na Pascal navazuje. Tento jazyk dovoluje velmi rychlé a bezpečné tvoření programového vybavení pro operační systém Windows, jeho nevýhodou je však nekompatibilita s operačním systémem Linux, kvůli čemuž jsem nemohl využít velkorysou nabídku Ústavu českého národního korpusu, abych pro své pokusy využil výpočetní kapacitu jejich serveru. Zmíněný software samozřejmě na požádání poskytnu každému, kdo by chtěl mé pokusy opakovat, a to nejen v rámci vědecké etiky, ale také jako podporu všem, kteří se vrhnou do bádání na tomto poli. Zároveň se budu snažit u všech měření uspokojivě popsat použitý algoritmus (aby tato měření mohli opakovat i lidé, kteří jsou zvyklí na jiný programovací jazyk) a text, na nějž byl aplikován.

Užité korpusy jsou většinou minimálně zpracované syrové autentické texty. Jediným zásahem bylo občasné odstranění zbytkové vokalizace u arabských textů, abychom sjednotili varianty jednoho typu, popřípadě odstranění interpunkce tam, kde by jen prodlužovala dobu výpočtu, a jiné drobné modifikace[23]. Texty procházely přísným výběrem a šetrným zpracováním, takže můžeme v souladu s Christianem Mairem (Mair, 2006: str. 355) použité korpusy zařadit do kategorie *small and tidy*. Skutečně, použité algoritmy dávaly uspokojivé výsledky i při zpracování

[23] Je dobrým zvykem korpusové lingvistiky různými způsoby zmenšovat variantnost textu (snižovat počet typů při zachování počtu tokenů), například pomocí lemmatizace. Snižuje se tím obvykle množství dat, které je potřeba k ověření nějaké hypotézy, popřípadě zmenšuje komputační náročnost úkolu. Na otázku, jestli lemmatizovaný text neztratí něco podstatného ze své struktury, si můžete odpovědět lehce tak, že z lemmatizovaného textu odstraníte všechny původní tokeny a pokusíte se ho číst. Z toho důvodu používám při práci originální texty a lemmatizaci ponechávám například lexikografům, kterým pomáhá v jednoznačném vyhledávání slov.

textů pod milion slov a osobní znalost obsahu většiny těchto textů mi pomáhala vyvarovat se hrubých chyb.

Výběr textů byl prováděn poněkud oportunisticky, vyhýbal jsem se těm, které nebyly jednotné, a nebo by mohly působit jakékoli obtíže; dával jsem přednost běžné narativní próze, přičemž základním kritériem byla dostupnost v elektronické podobě[24]. Samozřejmě zajímavé by bylo aplikovat použité algoritmy v „polních podmínkách" náhodně vybraných textů. Texty přeložené z jiných jazyků používám zejména tam, kde srovnávám výstupy algoritmu pro různé jazyky. V práci primárně zaměřené na moderní spisovnou arabštinu by bylo metodologicky čistější používat spíše překlady z arabštiny do jiných jazyků, nicméně jediným moderním arabským textem, který je po větách zarovnán se svým českým protějškem, je Kunderův *Valčík na rozloučenou*, a tak používám právě tento román.

V této souvislosti by bylo možno vytknout této studii[25], že se soustředí pouze na psaný jazyk. Tuto námitku není možné odbýt argumentem, že moderní spisovná arabština, která je v centru našeho zájmu, má své těžiště právě v psané podobě a mluví se jí jen v omezených situacích (některé sdělovací prostředky, vysoká literatura, prostředí arabských univerzit a zejména kateder humanitních oborů, nebo při islámských rituálních úkonech a vůbec při řešení islámské problematiky). Mluvený jazyk se od psaného liší v natolik důležitých věcech, že by bylo velmi odvážné vztáhnout na mluvený text zákony platné pro text psaný. Zásadní je například variantnost mluveného jazyka, která je nejen celkově vyšší, ale hlavně se nezastavuje na fonologické úrovni, jako je tomu v případě psaného jazyka, tedy vlastně žádné dva tokeny nejsou přesně stejné. Také dva tokeny, které bychom v psaném textu zcela jednoznačně přiřadili k různým typům si mohou být v mluveném textu podobnější, než dva tokeny, o kterých bychom v psaném textu řekli, že náleží ke stejnému typu. Ostré a jednoznačné hranice mezi tokeny i typy značně zjednodušují práci a umožňují formulovat jednoznačné definice, avšak otázka, jestli se nám tím neztrácí něco podstatného, je zcela na místě. Myslím, že komputační náročnost práce s velkými objemy mluvených textů by pro kvantitativní lingvisty neměla být omluvou a že bychom měli své hypotézy ověřovat i na mluveném textu, než je zobecníme pro celý jazyk.

Co se struktury práce, jejího členění a stylu týká, původní záměr byl, podobně jako v encyklopedii Laws in Quantitatie Linguistics[26], nejprve krátce uvést čtenáře do problému, pak

---

[24] Seznam použitých korpusů a jejich stručné charakteristiky naleznete na konci práce ve zvláštní kapitole Charakteristika použitých korpusů.

[25] A to zejména ze saussurovských pozic.

[26] http://lql.uni-trier.de/ (Altmann, Köhler, Vulanović 2006) Doporučuji kliknout na odkaz Random page. Tato stránka mi zároveň sloužila jako nejdůležitější heuristická báze.

uvést hypotézu, její odvození, metodu falzifikace a následně celou situaci ilustrovat změřenými daty. Takto by ovšem celá diplomová práce nenaplňovala požadovaný rozsah, neboť hypotézy kvantitativní lingvistiky se obvykle dají vyjádřit v několika větách – dokonce i ty nejdůležitější v historii byly velmi krátké[27]. Proto jsem přistoupil k trochu osobnějšímu způsobu psaní (to je také důvod, proč často opouštím autorský plurál a vyprávím v první osobě), který více méně biografickou formou popisuje cestu, jak jsem k řešení dospěl. Doufám, že i tyto informace budou pro čtenáře inspirativní a podnětné. K tomuto stylu jsem se inspiroval při četbě Wolframovy *A New Kind of Science* (Wolfram 2002), která mě uchvátila svou schopností podat i složitou myšlenku tak, že jí porozumí skutečně každý, a zároveň uměním pojednat o jednom jediném principu (ovšem přelomovém principu) na neuvěřitelných 800 stranách, aniž by se čtenář nudil.

[27] Vzpomeňme zásadní práci Alexeje Zipfa *Selected Studies of the Principle of Relative Frequency in Language* (Zipf 1923), která měla rozsah asi 120 stránek, z toho text zaujímal 24 stránek, přitom tento text – podle tehdejších i dnešních kritiků – nestál za mnoho (Wyllys 1981: str. 4). E. Prokosch dokonce na úvod své zdrcující recenze citoval verš z Goethova Fausta: *Ein grosser Aufwand, schmählich! ist vertan.* (Prokosch 1933). Z této knihy však do dnešních dnů přežila první aproximace závislosti frekvence typu na jeho pořadí ve slovníku (seřazeném od nejčetnějšího typu po nejméně četný) – dnes známá jako Zipfův zákon, nejdůležitější část tedy zabírala jeden řádek (nepočítáme-li dva grafy, které ilustrují použití tohoto vzorce).

1.6 Definice pojmů[28]

Jazyk Metoda, pomocí které může člověk sdělit svou myšlenku jinému člověku (materialisticky řečeno „popsat jinému člověku stav svého mozku“).

Text Výsledek snahy člověka o sdělení svých myšlenek jinému člověku. Konkrétní aplikace ↑jazyka.

Token Úsek ↑textu ohraničený z obou stran mezerami, interpunkčními znaménky nebo znakem konce odstavce, neobsahující na nehraničních pozicích žádnou mezeru, interpunkční znaménko ani znak konce odstavce.

Typ Forma daného ↑tokenu. Spojení *počet typů v textu* značí počet vzájemně různých ↑tokenů vyskytujících se v ↑textu.

Hapax Celým původním názvem *Hapax legomena*. ↑Typ, který je v daném ↑textu reprezentován právě jedním ↑tokenem.

Slovo Nabývá dvou významů: ↑typ a ↑token.

Věta Úsek ↑textu ohraničený z obou stran interpunkčními znaménky nebo znakem konce odstavce, neobsahující na nehraničních pozicích žádné interpunkční znaménko ani znak konce odstavce.

Četnost slova Přesněji *absolutní frekvence slova*. Četností ↑slova *x* v ↑textu *y* rozumíme počet ↑tokenů v textu *y*, které náleží ke stejnému typu jako token *x*.

Frekvenční zobrazení Též *zobrazení textu podle frekvence slov*: Zobrazení, které každému ↑tokenu v ↑textu přiřazuje číslo, jež vyjadřuje frekvenci ↑typu, ke kterému náleží, změřenou na daném ↑textu.

Délkové zobrazení Též *zobrazení podle délek tokenů*. Zobrazení, které každému ↑tokenu v ↑textu přiřazuje jeho délku v písmenech[29].

---

[28] Definice neřadím abecedně, ale tak, jak na sebe navazují.

[29] V arabských textech mezi písmena počítám jak souhlásky (ḥurūf), tak znaky pro samohlásky (ḥarakāt), neboť také ony pomáhají čtenáři při interpretaci významu. Pokud jsem vokalizaci z nějakého důvodu před pokusem z korpusu odstranil, pak na tuto skutečnost upozorňuji a důvod vysvětluji.

1.7 Typografická poznámka

Grafické uspořádání této práce není ideální, nicméně věřím, že jej čtenář shledá funkčním. Nejdůležitější věty jsem, navzdory starým dobrým typografickým pravidlům, vysázel tučným řezem, aby čtenáři neuniklo oněch několik málo exaktně formulovaných hypotéz, které bylo možno vyjádřit v několika větách, matematických vzorcích. Navzdory malému rozsahu, jejich vymýšlení zabralo mnoho času a energie, bloudění po slepých uličkách a mnohdy jich bylo dosaženo na cestě za zcela jiným cílem. Naopak vysvětlující text, domněnky, citáty autorit a ilustrující příběhy jsem zařadil do poznámek pod čarou (aby zbytečně nezdržovaly čtenáře, jenž se chce rychle dostat k jádru problému), které ovšem tímto způsobem mnohdy nabobtnaly nad únosnou mez.

Rovněž požadavek Filozofické fakulty UK, aby diplomová práce byla odevzdána v komerčním formátu Microsoft Wordu, nepříjemně omezuje výběr textového editoru na nepříliš kvalitní a drahé produkty firmy Microsoft (jsem zvyklý pracovat se sázecím systémem TeX) a snižuje grafickou úroveň této studie. Svým způsobem jsem na kvalitní typografii v této práci rezignoval, neboť sebemenší změny (například řezy písma, kontrola osamocených řádků apod.) mají za následek rozházení již napsaného textu, navíc některá – eufemisticky řečeno – nestandardní typografická nařízení (jedenapůlnásobné řádkování, úzké okraje stránky, dvanáctibodová velikost písma pro základní text…) jsou přímo požadovaná jako formát pro odevzdanou diplomovou práci[30].

Arabská slova a jména uvádím v přepisu, jaký používá *Zeitschrift der Deutschen Morgenländischen Gesellschaft*, což je ve středoevropských podmínkách tradiční standard. Pokud tuto práci čtete v elektronické podobě ve formátu Microsoft Wordu, je docela možné, že se vám některé znaky tohoto přepisu nezobrazily správně.

[30] Všechna nařízení týkající se grafické stránky diplomové práce jsou dostupné z http://www.ff.cuni.cz/FF-7296.html [11. 7. 2010].

2. Frekvence slov ve větě

2.1 Motivace

Prvním popudem k práci na tomto tématu byl seminář doc. Zemánka, kdy jsme při diskusi o staroarabské poezii narazili na fakt, že na **koncích arabských veršů[31] se objevují vzácná slova**. Toto tvrzení se odůvodňuje jednak tím, že arabští básníci nebo tradenti jejich veršů (*rāwī*, plurál *ruwāt*) verše v básních různým způsobem obměňovali, modernizovali při vývoji jazyka a přizpůsobovali jednotlivá slova publiku nebo události, takže verš mohl takto „žít" třeba i několik staletí. Jediné slovo, které zůstávalo zakonzervované, bylo slovo poslední, neboť sloužilo jako nositel rýmu. Dalším, možná podstatnějším důvodem pro vzácná slova na konci verše, se jevil fakt, že východní poetika pracuje s průběžným rýmem, takže najít po padesátém verši slovo, které se rýmuje s předchozími, sedí do metra, ještě nebylo použito a má alespoň přibližně odpovídající význam autorovu záměru, je složité. Básník pak sahá do hloubi paměti a vytahuje ze své slovní zásoby nejzapadlejší dialektizmy, archaizmy, nebo si potřebná slova aktivně vymýšlí.

Základní problém tohoto myšlenkového konstruktu (jinak logicky uspokojivého) je, že je postaven na dojmech. Přitom kvantifikace je v tomto případě jednoduchá.

2.2 Metoda měření

Provedeme **zobrazení textu podle frekvence slov**: Mějme korpus staroarabských veršů. **Každému slovu přiřaďme jeho četnost v tomto korpusu**[32].

Poté roztřídíme verše podle jejich délky. Dostaneme tak data, jako například v této tabulce[33]:

Tabulka číslo 1:

[31] Správněji bychom měli říkat dvojverší.

[32] Precizněji řečeno, každému tokenu přiřaďme četnost typu, ke kterému náleží.

[33] Výňatek z frekvenčního zobrazení korpusu KOMPLET (viz kapitolu Charakteristika použitých korpusů). Kvůli nedostatečné velikosti se mezi hapaxy řadí i poměrně častá slova, takže vypadají, že jsou vzácnější, než ve skutečnosti. Čím větší je korpus, tím přesnější je ohodnocení frekvence slov.

Původní text (zprava doleva)	Délka	Zobrazení (zleva doprava)
ديار لسلمى عافيات بذي خال ألح عليها كل أسحم هطال	10	2 1 1 9 2 1 6 27 2 2
وتحسب سلمى لا تزال ترى طلا من الوحش أو بيضا بميثاء محلال	12	2 4 52 2 14 1 158 2 40 1 1 1
وتحسب سلمى لا نزال كعهدنا بوادي الخزامى أو على رس أوعال	11	2 4 52 1 1 1 1 40 84 1 1
ليالي سلمى إذ تريك منصبا وجيدا كجيد الرئم ليس بمعطال	10	1 4 14 2 1 1 2 2 15 1
ألا زعمت بسبابة اليوم أنني كبرت وأن لا يحسن اللهو أمثالي	11	14 1 1 4 5 1 4 52 1 1 2
كذبت لقد أصبى على المرء عرسه وأمنع عرسي أن يزن بها الخالي	12	1 3 1 84 4 1 1 1 26 1 25 2

Pak zprůměrujme hodnoty na jednotlivých pozicích pro verše o stejných počtech slov. Například průměrná řada pro verše o deseti slovech by v této tabulce byla:

Tabulka číslo 2:

1,5	2,5	7,5	5,5	1,5	1	4	14,5	8,5	1,5

Z dat v první tabulce je vidět, že čísla na jednotlivých pozicích mají značné rozpětí a že tedy průměrem nahrazujeme skupinu čísel, jež se k průměru málokdy přiblíží (průměrná odchylka činí u tohoto korpusu zhruba 130 % průměru). Je to tím, že slova se řídí Zipfovou distribucí a nikoliv normálním rozdělením. Také proto postrádá smysl určovat směrodatnou odchylku a používat statistické nástroje na ní založené. Je tedy potřebné zdůraznit, že zákonitosti, které bude možné v těchto datech nalézt, se budou projevovat teprve ve velké mase textu[34].

Otázkou tedy je, jak velkou statistickou významnost můžeme naměřeným hodnotám (jako jsou ve druhé tabulce) přikládat. Pokud bychom data ze druhé tabulky naměřili na korpusu se zpřeházenými tokeny, budou hodnoty na všech pozicích oscilovat kolem jedné hodnoty, a to podle normálního rozdělení. Můžeme tedy přibližně změřit, jak velký je variační koeficient pro korpusy o určité velikosti, a podle toho odhadnout, jak velké rozdíly v hodnotách jednotlivých pozic jsou dány náhodně a jak velké musí být, abychom mohli mluvit o struktuře textu.

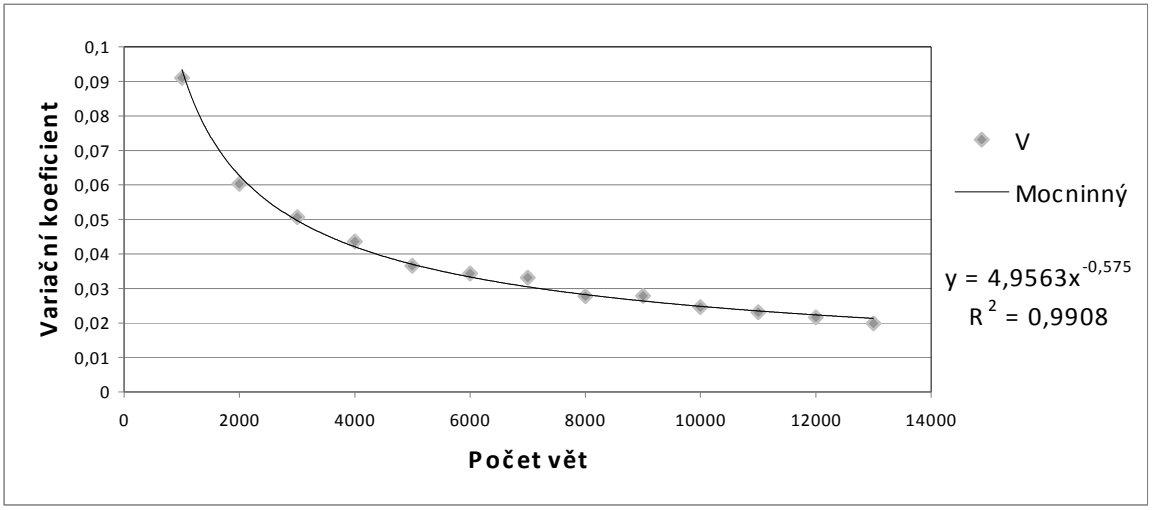
Následující graf nám ukáže závislost variačního koeficientu[35] na počtu zprůměrovaných celků (které si můžeme představit jako verše nebo věty) na náhodně zpřeházeném korpusu JM[36]:

---

[34] Nejsem si jistý, jestli každý čtenář této práce je srozuměný s myšlenkou existence stochastických zákonů, proto raději zopakuji, že říkám-li např. „průměrné poslední slovo ve větě má nižší než průměrnou frekvenci“, nemyslím tím, že poslední slovo v *každé* větě má nižší frekvenci, než je průměr pro tuto větu. Toto téma rozebírám podrobněji v kapitole 1.3.3.

[35] Variačním koeficientem datové řady rozumíme standardní odchylku datové řady dělenou jejím aritmetickým průměrem.

[36] Korektněji bychom měli říkat „korpusu s náhodně transponovanými tokeny“. Více o arabském korpusu JM najdete v kapitole 9.



Vidíme, že již kolem 1000 celků se dostáváme na vcelku uspokojivý variační koeficient 9 %, z charakteru křivky mocninného vztahu vyplývá, že po 10000 celků se variační koeficient snižuje jen pozvolna. Podobný graf si vytvoříme pro všechny korpusy, se kterými budeme pracovat, a podle toho budeme u každého měření dodávat předpokládaný variační koeficient.

2.3 Aplikace této metody

Nad soustavou dat, která vznikne popsaným měřením frekvenčního zobrazení, si můžeme klást tyto otázky:

1) Liší se průměrná frekvence slov podle jejich použití ve verši? Má poslední slovo skutečně nejmenší průměrnou frekvenci?
2) Liší se z tohoto hlediska různě dlouhé verše nebo verše s různým metrem?

Na zodpovězení první otázky je vhodné podívat se na grafy, které z dat vzešly[37].

[37] Tyto grafy pocházejí z dat naměřených na verších o devíti, deseti a jedenácti slovech od různých staroarabských básníků, zejména ze sbírky *Muʿallaqāt* (korpus KOMPLET, viz kapitolu Charakteristika použitých korpusů). V korpusu jsou zastoupena různá metra, celkově 224 veršů o devíti slovech, 212 veršů o deseti slovech a 142 veršů o jedenácti slovech. Nedostatečná velikost korpusu se podepsala na kvalitě dat (u korpusu o 200 větách činí na náhodně zpřeházeném textu bez struktury variační koeficient průměrných frekvencí slov přibližně 23 %), takže podobnost mezi grafy není tak velká jako v následující části práce, která rozebírá tentýž jev na mnohem větším korpusu moderní arabské prózy.

Ukazuje se na nich, že **původní předpoklad platí, zároveň také vykazují určité rysy složitější struktury**. Tato „velbloudí křivka[38]“ se vytvořila pro soubory veršů s různou délkou i metrem, avšak fatální nedostatek staroarabské poezie ve snadno zpracovatelném elektronickém formátu mě přinutil tento výzkum předčasně ukončit.

Proto jsem se tehdy nedostal ke zodpovězení třetí a nejdůležitější otázky, ke které jsem se vrátil až nyní, při řešení diplomové práce:

Je tato struktura typická pouze pro staroarabskou poezii, nebo se s ní můžeme setkat v jakémkoli arabském textu?

Arabský verš je totiž samostatná významová jednotka, přesahy se ve staroarabské poezii téměř nevyskytují[39], značně tedy připomíná větu (klauzi) a téměř vždy se kryje s jejími hranicemi. Aplikujeme proto zmíněný algoritmus na moderní arabský text[40] a místo veršů analyzujeme oznamovací věty[41]. Nyní nechme hovořit data[42]:

---

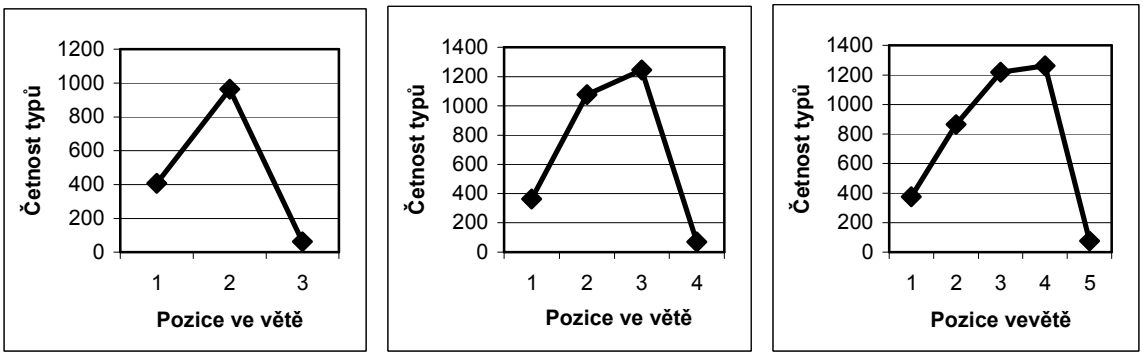
[38] Na posledním grafu s jedním hrbem a hlavou. Toto možná poněkud nemístně metaforické pojmenování (které vzniklo spontánně na kolejích při diskuzi nad daty s kolegou Přemyslem Kubátem, znalcem staroarabské poetiky a předislámské arabštiny, snad analogicky k hokejkovému grafu Michaela Manna) odráží také hru s kořenem arabského slova označujícího větu (*ǧumla*) a slova pro velbloudа (*ǧamal*).

[39] Verše mohou v závislosti na recitátorovi měnit pořadí nebo dokonce putovat z básně do básně. Typický je proto arabský termín pro verš (*bayt*), který může zároveň znamenat *dům*, *stan*, *domácnost* nebo *nukleární rodinu* (manžele a děti).

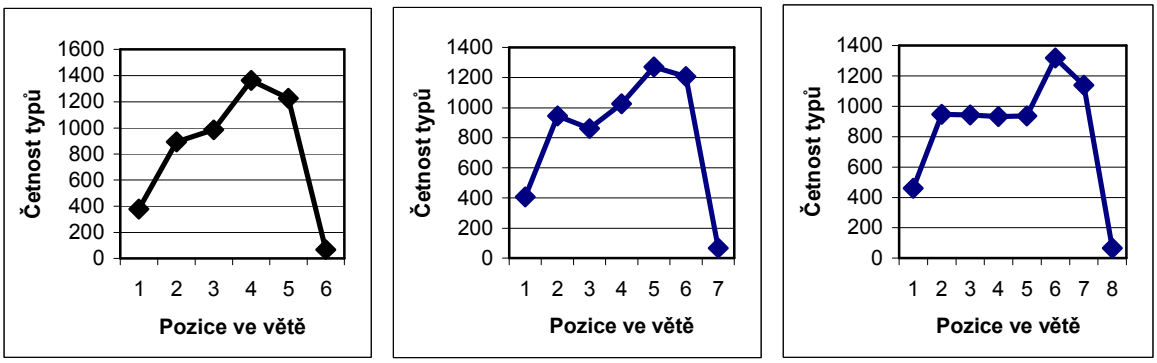
[40] Kolekci několika desítek moderních povídek různých arabských autorů, korpus AWU (viz kapitolu Charakteristika použitých korpusů).

[41] Množinu všech úseků textu, které začínají tečkou, vykřičníkem, otazníkem, čárkou, středníkem nebo dvojtečkou a končí tečkou, dvojtečkou, čárkou nebo středníkem, přičemž žádné z těchto interpunkčních znamének není použito vevnitř úseku.

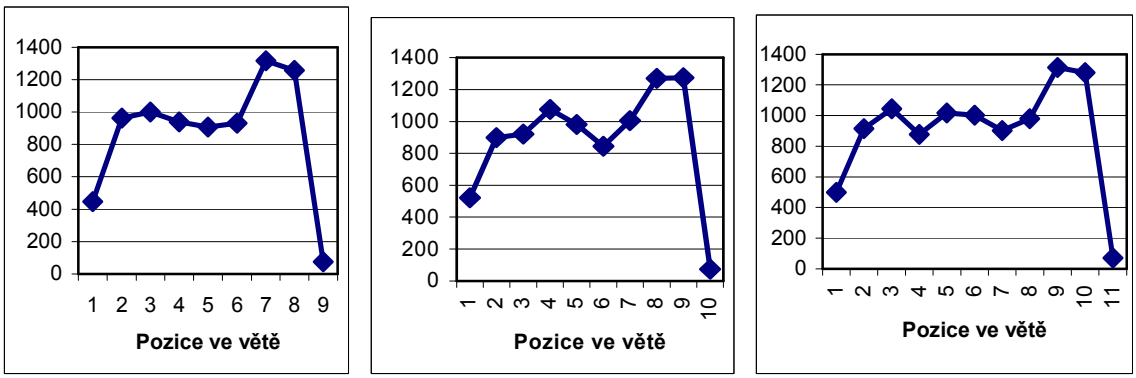
[42] Jedná se o grafy získané ze syrových dat, která nebyla nijak statisticky upravována.



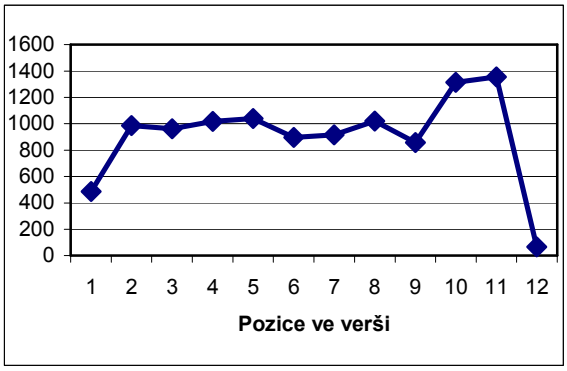
Ve větách o třech až pěti slovech vytváří graf jednu vlnu. Počet zprůměrovaných vět je v prvním případě 7300, ve druhém 6800 a ve třetím 5900, očekáváme variační koeficient nad 3,5 %. Je tedy bez debaty, že nízká frekvence posledního tokenu ve větě odpovídá situaci v poezii.

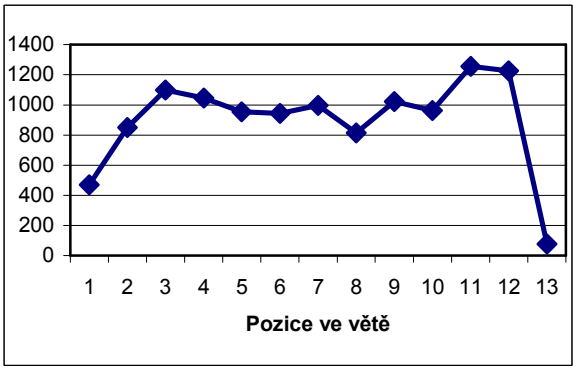


Od šestislovných vět získávají grafy typický „velbloudovitý“ tvar, který je pro arabštinu charakteristický, objevuje se i v textech psaných starou arabštinou, avšak u evropských jazyků má poněkud odlišné rysy. O testech provedených na korpusech středověkých arabských textů (JM a CHALDUN) bude řeč dále, stejně jako o testech provedených na korpusech západoevropských jazyků. Počet vět se pohybuje mezi pěti a třemi tisíci, přibližná hodnota variačního koeficientu tedy nepřesahuje 5 %.



Na větách o devíti, deseti a jedenácti slovech je již patrné, že náhodné odchylky tvar křivky ovlivňují. Počet vět, na kterých byly grafy získány, jsou nízké (cca 2500 – 1500), takže variační koeficient stoupá na 6 – 7,5 %.

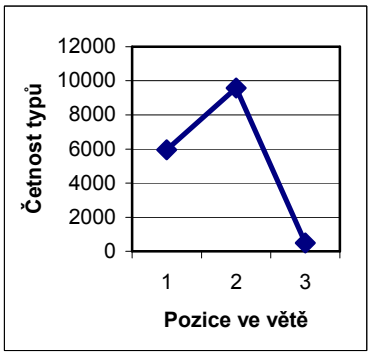


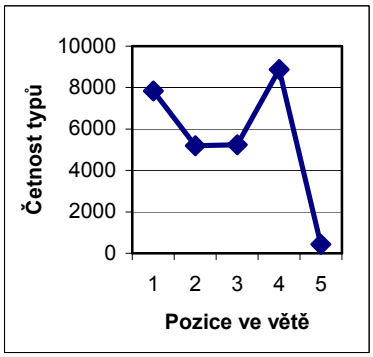


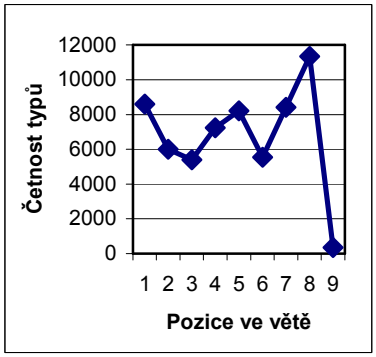
Věty o dvanácti a třinácti slovech vytvářejí podobný obrazec jako předchozí grafy, avšak data již nejsou tak dobře čitelná – například vět o třinácti slovech byl v našem korpusu jeden tisíc, což odpovídá variačnímu koeficientu 9 %

Všimněme si rysů, které jsou všem těmto grafům společné: typ na první pozici má četnost okolo 400 tokenů, průměrný typ na předposlední pozici okolo 1200 tokenů a typ na poslední pozici okolo 70 tokenů, a to nezávisle na délce věty. Mezi třetím tokenem a třetím tokenem odzadu vzniká jakési plateau, kde se průměrná frekvence pohybuje okolo 950 tokenů a odchylky od této hodnoty jsou v rámci variačního koeficientu.

Výsledek mě natolik nadchl, že jsem adekvátní algoritmy aplikoval i na anglický text[43]. Nebudeme se již zdržovat uváděním všech grafů, neboť tato práce je primárně zaměřena na arabštinu, a vybereme jen několik ilustrativních výsledků:



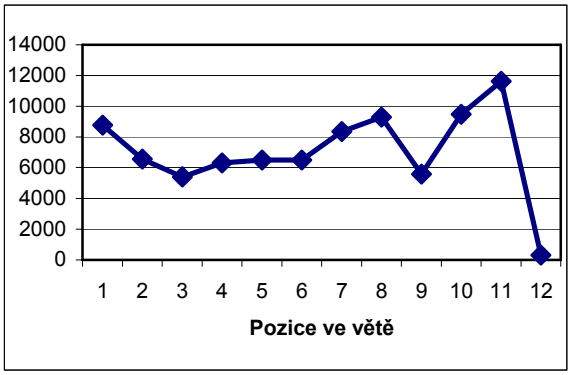


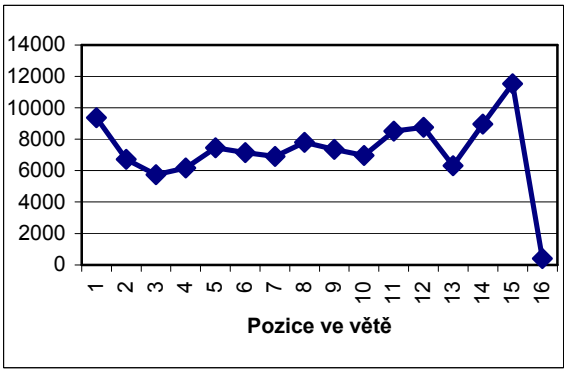


Graf pro věty o třech slovech (z 5400 vět, variační koeficient je asi 2,3 %) se velmi podobá svému arabskému protějšku, avšak u dalších grafů (z 5900 a z 5700 vět, variační koeficient je asi 2,2 %) je jasně vidět, že typická anglická klauze začíná nadprůměrně četným typem. Také pro

[43] Soubor 3 románů J. F. Coopera (korpus COOPER), malá písmena byla nahrazena velkými, aby se na počátku vět nevyskytovalo nepřiměřeně mnoho vzácných slov (slov s velkým počátečním písmenem).

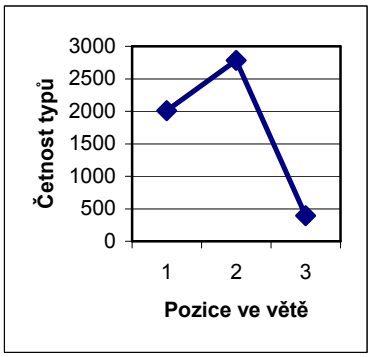
angličtinu platí jednotná průměrná četnost typu prvního slova (cca 8500 výskytů) i předposledního (asi 11000 výskytů) a posledního (kolem 350 výskytů).

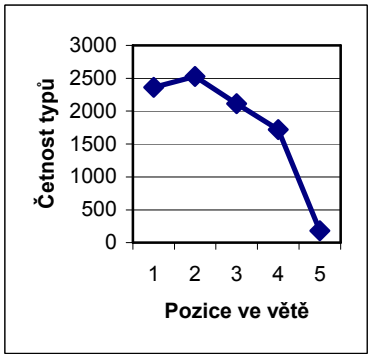


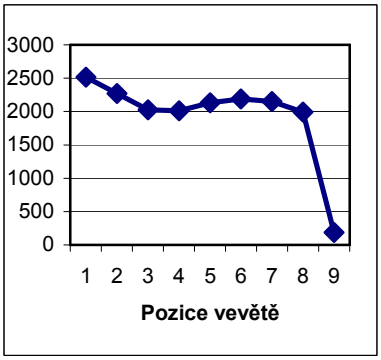


Charakteristická struktura se netýká jen prvních dvou a posledních dvou slov, jak vidíme na grafech o 9, 12 a 16 slovech, také frekvence tokenů, které jsou 3. a 4. odpředu i odzadu zůstávají stejné. Věty o více slovech mají ze statistických důvodů (je jich méně než 1000 a variační koeficient stoupá nad 10 %) strukturu zamlženou, avšak i v datech pro věty o 18 slovech můžeme najít náznak odpovídající struktury. Takto dlouhé věty ale zcela jistě nejsou jednotlivé klauze, ale spíše souvětí, která nebyla v textu rozdělena interpunkčním znaménkem.

Velký rozdíl mezi předposledním a posledním tokenem, jakož i další systémové podobnosti s výsledky získaných na arabštině, mě povzbuzuje k dalším pokusům na tomto poli. Pro zajímavost si uveďme výsledky, jichž dosáhl zmíněný algoritmus na českém korpusu[44]:







Vět o třech slovech bylo v korpusu 9400 (variační koeficient je asi 2,5 %), vět o 5 slovech bylo 8000 (variační koeficient je asi 2,7 %), vět o 9 slovech bylo 3000 (variační koeficient je asi 5 %), dále počty slov prudce klesají, neboť český pravopis má tendenci oddělovat jednotlivé klauze čárkami (a ty počítáme jako hranice vět). Struktura věty o třech slovech je opět podobná arabštině i angličtině. Co také zůstává, je základní zjištění, se kterým jsme k tomuto testu přistupovali, a

[44] Jedná se o sbírku několika beletristických děl Karla Čapka (divadelní hry jsem pro uchování konzistence s předchozími pokusy nezařadil). Malá písmena byla nahrazena velkými, aby se na počátku vět nevyskytovalo nepřiměřeně mnoho vzácných slov (slov s velkým počátečním písmenem).

totiž, že na konci věty se hromadí relativně vzácná slova (v tomto případě mají typy na posledních pozicích průměr okolo 200 výskytů v textu, předposlední slovo pak zhruba desetinásobek).

2.4 Interpretace

Hypotéza číslo 1: **Poslední token ve významovém celku průměrně náleží k typu, který má v textu nižší než průměrnou frekvenci.**

Tato hypotéza je vyvratitelná pomocí frekvenčního zobrazení (za významové celky považujeme věty, jak jsme si je definovali). Podívejme se, jak platí na různých jazycích[45]:

Korpus	**Poslední token[46]**	**Předposlední token[47]**	**Průměrný token[48]**
AWU	69,19	1188,10	737,35
COOPER	396,89	10247,29	6167,13
MLOCI	228,95	2044,49	228,96
NAHODNY	471,24	454,62	485,86
JM	403,66	3055,23	1993,58
CHALDUN	789,57	7492,68	4458,51
HUGO	1190,60	14535,44	9613,09
ZOLA	926,24	12655,04	8258,42

Ve všech případech, kromě korpusu NAHODNY, který byl zařazen jako kontrolní vzorek pro detekci hrubých chyb, je předposlední slovo ve větě více než pětkrát četnější než slovo průměrné a tento rozdíl je ještě větší, pokud porovnáváme poslední a předposlední slovo.

Teze, že na konci významových celků se vyskytují vzácná slova, pro staroarabskou poezii sice platí, ovšem tento popis nám neříká celou pravdu, a totiž že tento stav není privilegiem staroarabské poezie, ale že ho najdeme v běžném arabském textu, ať už středověkém (korpusy JM a CHALDUN), nebo moderním (korpus AWU), a dokonce i v jiných jazycích (v angličtině –

---

[45] Popisy jednotlivých korpusů naleznete v kapitole Charakteristika použitých korpusů, korpus NAHODNY vznikl náhodnou transpozicí tokenů v korpusu AWU (zkráceném) a následným náhodným rozdělením do „vět".

[46] Průměrná absolutní frekvence typů, k nimž náleží poslední token ve větách o třech až dvanácti slovech.

[47] Průměrná absolutní frekvence typů, k nimž náleží předposlední token ve větách o třech až dvanácti slovech.

[48] Průměrná absolutní frekvence všech typů v textu (započteny tolikrát, kolikrát se vyskytují ve větách o třech až dvanácti slovech).

korpus COOPER, v češtině – korpus MLOCI a ve francouzštině – korpusy HUGO a ZOLA). Naopak na korpusu s náhodně transponovanými tokeny (korpus NAHODNY) tento vztah neplatí, čili není výsledkem hrubých chyb při měření.

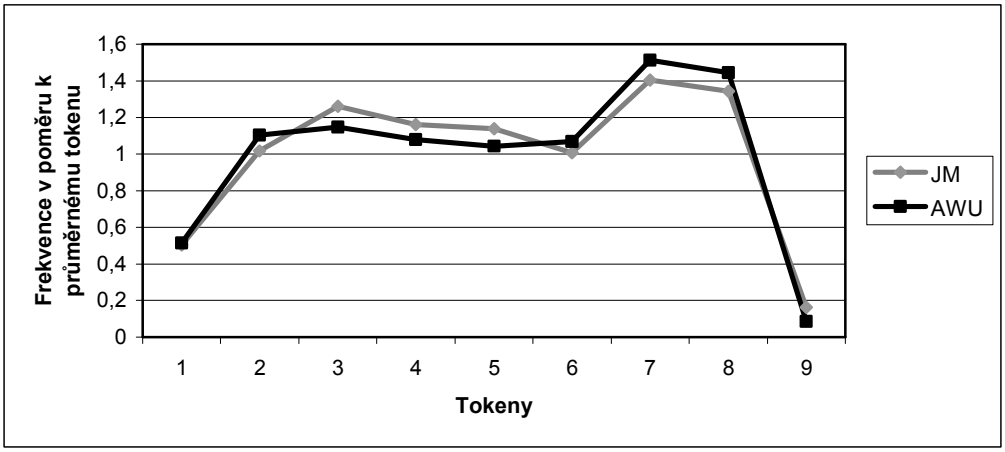
Prozatím budeme opatrní a nebudeme tvrdit, že platí ve všech jazycích, i když se nám výjimky dosud vyhýbaly a vybrané evropské jazyky spolu s arabštinou tvoří typologicky poměrně různorodou směs. Dokážu si nicméně představit, že v jazycích, které mají postponovaná častější slova (například ukazovací zájmena, členy, slova ve funkci našich předložek…), bude tento vztah méně jasný, nebo nebude platit vůbec. Korpus takového jazyka ovšem nemám k dispozici.

Hypotéza číslo 2: **Průměrné věty v textech napsaných stejnými jazyky mají ve frekvenčním zobrazení podobnou strukturu**.

Hypotéza číslo 3: **Průměrné věty v textech napsaných různými jazyky mají ve frekvenčním zobrazení podobnou strukturu.**

Podíváme se na tyto dvě hypotézy současně.

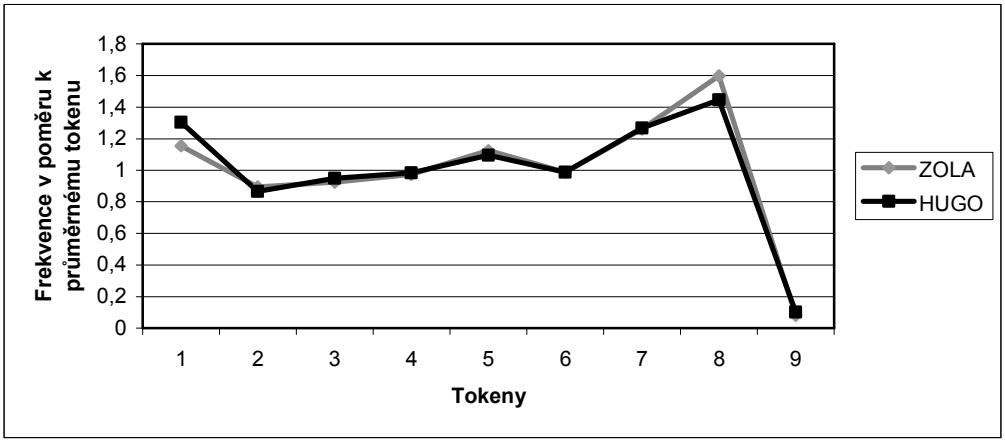
Napřed uvedeme grafy, které budou ilustrovat druhou hypotézu.



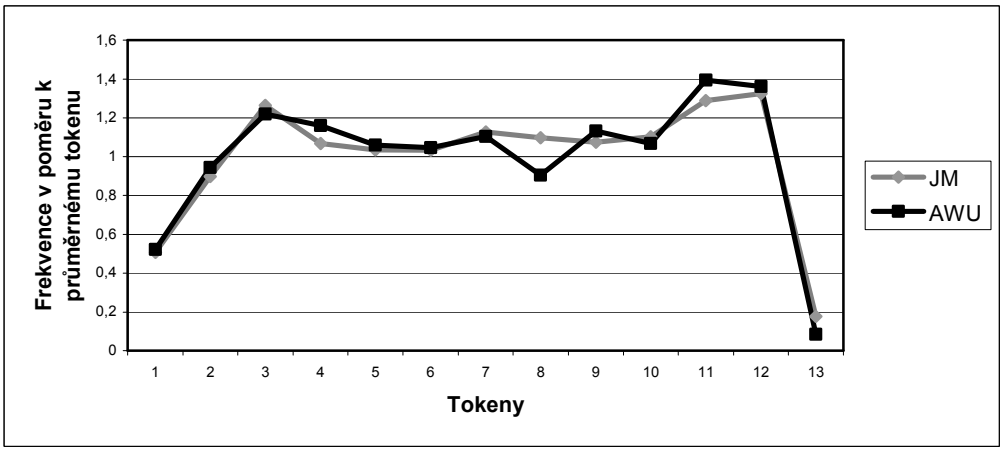
Světlejší křivka byla naměřena na devítislovných větách z al-Masʿūdīho a al-Ǧāḥiẓových vědeckých kompendií, tmavší na sbírce moderních arabských románů a povídek[49]. Přestože tyto texty dělí tisíc let, na grafu je dobře vidět, že základní struktura věty se nezměnila. Což teprve,

[49] Variační koeficient pro devítislovné a dvanáctislovné věty u těchto korpusů uvádím pro přehlednost níže v tabulce.

když porovnáme věty současníků – šedá linka ilustruje data získaná na velkých románech Émila Zoly, černá zahrnuje data z románů Viktora Huga a Alexandra Dumase:

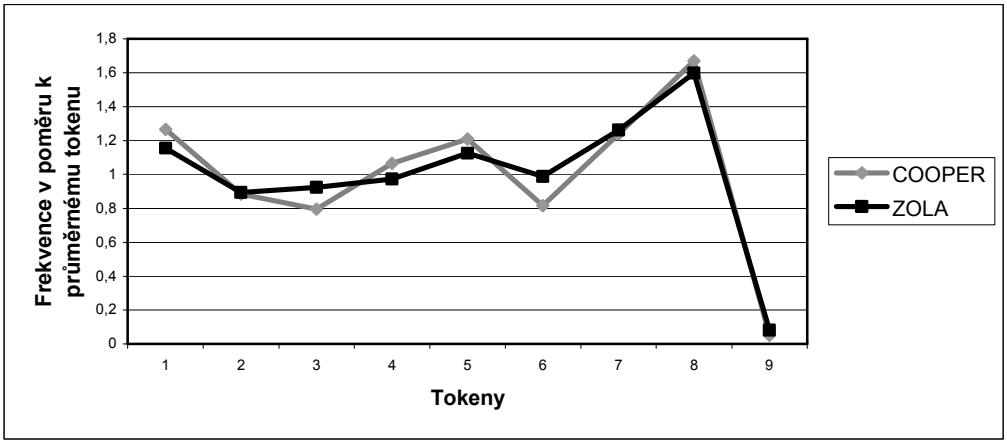


Vidíme, že křivky se nejen podobají, ale spíše se překrývají tak, že často není jedna přes druhou vidět. Zkusíme, jestli se podobná struktura týká pouze relativně krátkých vět, nebo jestli postihuje i ty delší:

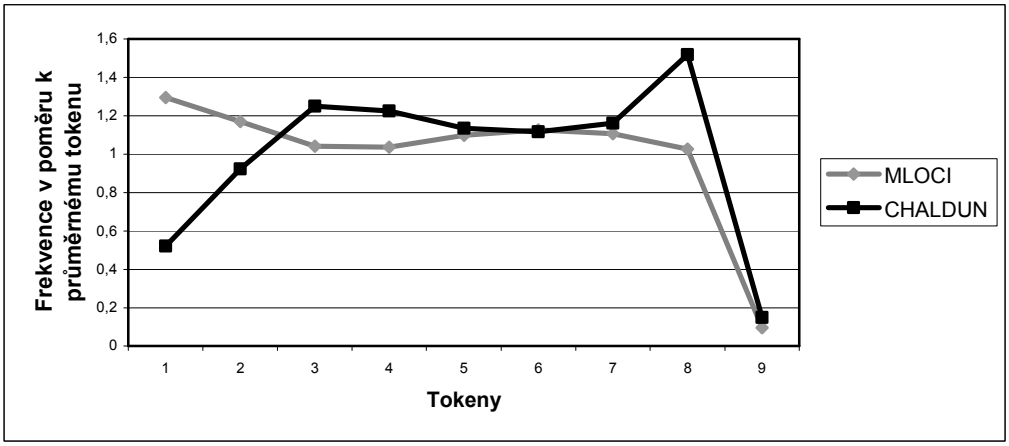


Pozorujeme, že i na delších větách je podobnost celkem uspokojivá, i když ne tak přesvědčivá jako u kratších celků.

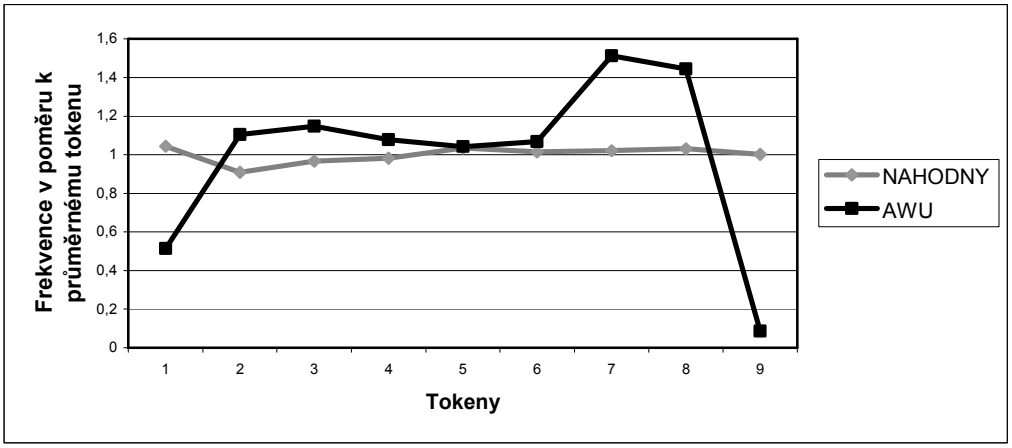
Myslím, že se shodnu se čtenářem na předběžném dojmu, že texty psané stejným jazykem vykazují podobnou strukturu věty. Nyní si zkusme stejně subjektivním způsobem srovnat výsledky naměřené na textech různých jazyků:



První na řadě je soubor románů J. F. Coopera, který srovnáváme se sbírkou jeho francouzského mladšího kolegy Émila Zoly. I zde se křivky na první pohled podobají. Méně přesvědčivě však dopadne srovnání textu českého (výběr Čapkových románů, korpus MLOCI) a arabského (korpus CHALDUN):



Mohli bychom prozatím uzavřít konstatováním, že jsou-li jazyky strukturně podobné (například areálně spřízněná angličtina s francouzštinou), grafy vypadají podobně; když jsou si jazyky vzdálené (jako například arabština s češtinou), podobnost stále existuje, ale musíme zapojit více fantazie, abychom ji viděli. Pořád je ale větší, než když srovnáme výsledky přirozeného jazyka s hodnotami naměřenými na „textu“ s náhodně zpřeházenými tokeny, náhodně rozdělenými do celků po devíti tokenech:



Není však možné spokojit se pouze se subjektivními dojmy z několika málo grafů, pro vyčíslení podobnosti křivek nám statistika nabízí (mimo jiné) Pearsonův korelační koeficient a determinační koeficient[50]. Nejprve shromáždíme zprůměrované věty o devíti slovech z různých korpusů (podobně jako u předchozí hypotézy jsme vedle korpusů popsaných v kapitole Charakteristika použitých korpusů zařadili kontrolní korpus NAHODNY). Pro každou dvojici vět vypočítáme Pearsonův korelační koefiocient (R):

R pro věty o devíti slovech:

	AWU	**COOPER**	**MLOCI**	**NAHODNY**	**JM**	**CHALDUN**	**HUGO**	**ZOLA**
AWU	**1,0000**	0,7002	0,6375	*-0,0682*	**0,9799**	**0,9317**	0,7194	0,7906
COOPER		**1,0000**	0,7583	*0,3789*	0,6607	0,6953	**0,9651**	**0,9772**
MLOCI			**1,0000**	*0,0260*	0,6103	0,5940	0,8543	0,7987
NAHODNY				**1,0000**	*-0,0801*	*-0,0063*	*0,3770*	*0,3428*
JM					**1,0000**	**0,9583**	0,6763	0,7451
CHALDUN						**1,0000**	0,6936	0,7761
HUGO							**1,0000**	**0,9824**
ZOLA								**1,0000**

Stejným způsobem pro každou dvojici vět vypočteme determinační koeficient (R^2):

[50] Pearsonův korelační koeficient vyjadřuje vztah mezi datovými řadami. Nabývá hodnot od -1 do 1, -1 značí úplnou nepřímou úměru, 0 žádnou souvislost mezi daty, 1 úplnou přímou úměru. Determinační koeficient je druhou mocninou Pearsonova korelačního koeficientu a vyjadřuje, jak moc je možné vysvětlit jednu datovou řadu druhou datovou řadou. Tedy například pokud je determinační koeficient pro dva grafy roven 0,9, pak můžeme říci, že křivka v jednom grafu je vysvětlitelná z 90 % křivkou druhého grafu. Základní statistická poučka praví, že pozitivní korelace nezakládá příčinný vztah (o tom více v kapitole Epistemologická východiska).

	AWU	COOPER	MLOCI	NAHODNY	JM	CHALDUN	HUGO	ZOLA
AWU	**1,0000**	*0,4903*	*0,4064*	*0,0047*	**0,9601**	0,8680	0,5175	0,6250
COOPER		**1,0000**	0,5751	*0,1436*	*0,4365*	*0,4835*	**0,9315**	**0,9549**
MLOCI			**1,0000**	*0,0007*	*0,3724*	*0,3529*	0,7298	0,6380
NAHODNY				**1,0000**	*0,0064*	*0,0000*	*0,1422*	*0,1175*
JM					**1,0000**	**0,9183**	*0,4574*	0,5551
CHALDUN						**1,0000**	*0,4811*	0,6024
HUGO							**1,0000**	**0,9651**
ZOLA								**1,0000**

Abychom věděli, jak velkou váhu můžeme těmto korelacím přisoudit, určeme si, jak velký je variační koeficient jednotlivých průměrných vět.

Korpus	**Počet vět o devíti slovech**	**Variační koeficient (%)**
AWU	2372	6
COOPER	4508	4,2
MLOCI	2924	5
NAHODNY	1854	4
JM	4966	3,6
CHALDUN	11463	2,5
HUGO	10917	2,7
ZOLA	10900	2,8

Devět prvků je pro použití korelačního a determinačního koeficientu poměrně málo. Než se pustíme do interpretace předchozích tabulek, zvýšíme počet prvků na dvanáct (použitím vět o dvanácti slovech). Ovšem za cenu zvýšení variačního koeficientu.

R pro věty o 12 slovech:

	AWU	COOPER	MLOCI	NAHODNY	JM	CHALDUN	HUGO	ZOLA
AWU	**1,0000**	0,7497	0,5898	*-0,5405*	**0,9472**	**0,9358**	0,7455	0,8198
COOPER		**1,0000**	0,7390	*-0,4890*	0,6355	0,6548	**0,9518**	**0,9621**
MLOCI			**1,0000**	*-0,5263*	0,5522	0,5302	0,7807	0,7404
NAHODNY				**1,0000**	*-0,4192*	*-0,4235*	*-0,4537*	*-0,4331*
JM					**1,0000**	**0,9588**	0,6640	0,7381
CHALDUN						**1,0000**	0,6951	0,7638
HUGO							**1,0000**	**0,9852**
ZOLA								**1,0000**

Stejným způsobem pro každou dvojici vět vypočteme determinační koeficient (R^2):

	AWU	COOPER	MLOCI	NAHODNY	JM	CHALDUN	HUGO	ZOLA
AWU	**1,0000**	0,5620	*0,3478*	*0,2922*	**0,8972**	**0,8756**	0,5558	0,6720
COOPER		**1,0000**	0,5461	*0,2391*	*0,4038*	*0,4288*	**0,9060**	**0,9256**
MLOCI			**1,0000**	*0,2770*	*0,3049*	*0,2811*	0,6095	0,5483
NAHODNY				**1,0000**	*0,1757*	*0,1794*	*0,2059*	*0,1876*
JM					**1,0000**	**0,9193**	*0,4408*	0,5448
CHALDUN						**1,0000**	*0,4832*	0,5834
HUGO							**1,0000**	**0,9707**
ZOLA								**1,0000**

Korpus	Počet vět o dvanácti slovech	Variační koeficient (%)
AWU	1204	8
COOPER	2770	3,5
MLOCI	1302	7,8
NAHODNY	1091	9,4
JM	2485	5,5
CHALDUN	6554	3,5
HUGO	6206	3,4
ZOLA	4418	3,8

Jak pro R tak pro R^2 zvýrazňujeme pro přehlednost dvojice, které mají výsledek nad 90 % tučně a naopak výsledek pod 50 % kurzívou. Pro devítislovné i dvanáctislovné věty se ukazuje, že nejlépe korelují hodnoty pro korpusy psané týmž (nebo alespoň podobným) jazykem, zejména vyčnívá arabská trojice AWU, CHALDUN a JM a francouzská dvojice ZOLA – HUGO (98 % korelace devítislovných vět!), ke které se s notným odstupem připojuje anglicky psaný COOPER. Zcela stranou stojí český korpus MLOCI, jehož průměrné devítislovné věty nejlépe korelují s průměrnými devítislovnými větami Émila Zoly a všechny ostatní hodnoty leží v intervalu od 50 do 80 %, determinační koeficient je pak s arabským textem na úrovni náhodného textu.

Samotný náhodný text (korpus NAHODNY) se chová přesně podle očekávání – vykazuje nízkou zápornou korelaci, nebo velmi nízkou kladnou korelaci, často nulovou.

Pohledem na tabulku můžeme zhodnotit, že tyto dvě hypotézy jsme tímto testem nevyvrátili.

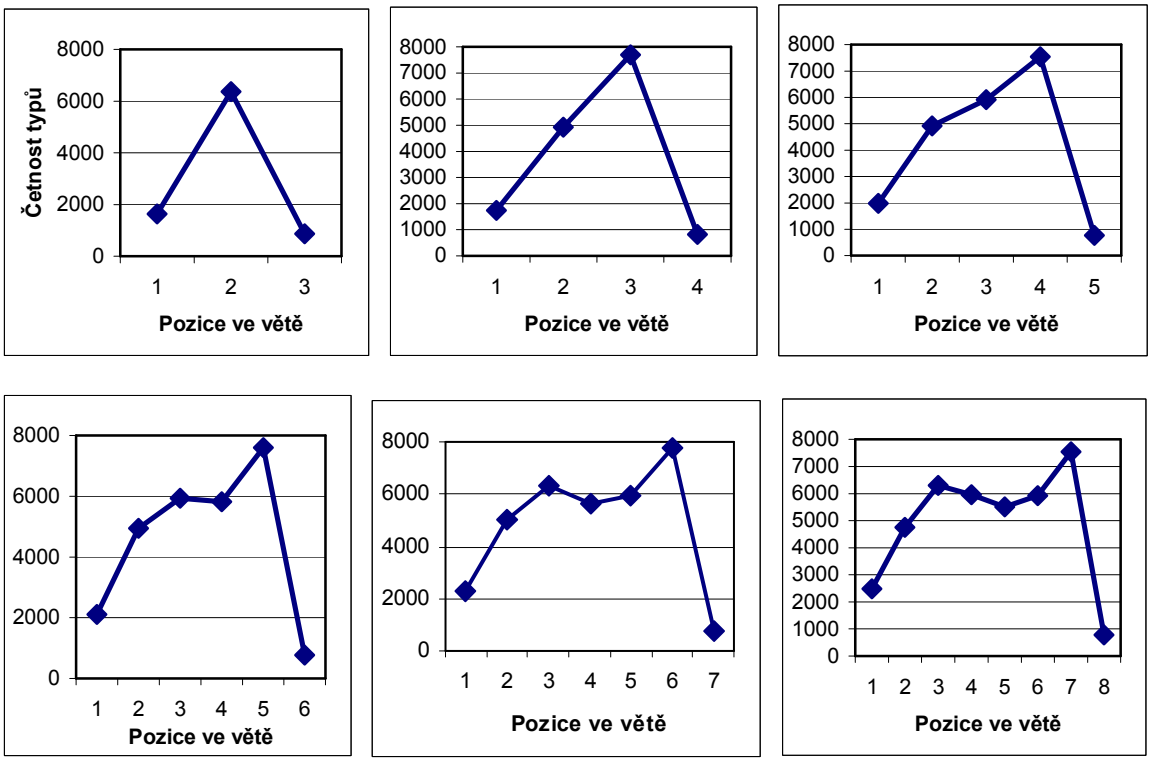
Hypotéza číslo 4: **V rámci jednoho textu mají průměrné věty o různé délce ve frekvenčním zobrazení podobnou strukturu.**

Pearsonův korelační koeficient a od něj odvozený determinační koeficient je definován pouze pro srovnávání dvou množin o stejném počtu prvků. Pokud srovnáváme graf pro věty o n slovech s grafem pro věty o n+1 slovech, máme dvě možnosti, jak se s tím vypořádat: rozpočítat n+1 hodnot na n hodnot[51], nebo nějaký token v datové řadě pro věty o n+1 tokenech vypustit. Podívejme se, co je výhodnější.

Napřed se necháme inspirovat pohledem na grafy[52]:

[51] Například přiřazením hodnoty nejbližšího datového bodu, více o tomto postupu v kapitole 5.3. Hodí se spíše pro datové řady o mnoha prvcích.

[52] Byly získány na jeden a půl milionovém arabském korpusu CHALDUN. První graf vznikl zprůměrováním frekvenčního zobrazení 14400 vět (variační koeficient je 1,9 %), druhý a třetí 17000 vět (variační koeficient je přibližně 1,7) čtvrtý 16000 vět (variační koeficient je asi 1,75) pátý 14 500 vět o sedmi slovech (variační koeficient je 1,9 %), šestý 13000 vět (variační koeficient je 2 %) o osmi slovech.



Jsme svědky podivného jevu. Každý graf je tvořen daty, která byla naměřena na úplně jiných větách, přesto to vypadá, jako by se jednalo o vývojovou řadu:

5. graf jako by vznikl vypuštěním 4. tokenu v 6. grafu.
4. graf jako by vznikl vypuštěním 4. tokenu v 5. grafu.
3. graf jako by vznikl vypuštěním 3. nebo 4. tokenu ve 4. grafu.
2. graf jako by vznikl vypuštěním 3. tokenu v 3. grafu.
1. graf jako by vznikl vypuštěním 2. tokenu ve 2. grafu.

Vidíme celkem jednoduchou řadu: čtvrtý z osmi, čtvrtý ze sedmi, třetí z šesti, třetí z pěti, druhý ze čtyř – tedy prostřední z lichého počtu, *n*/2 -tý ze sudého počtu[53].

Formulujme tento algoritmus precizněji: „Z datové řady o *n* prvcích vypusť prvek, který má pořadí n/2, přičemž pořadí se zaokrouhluje na celá čísla.“ Nazvěme si tento algoritmus jako **vypuštění prostředního tokenu** .

Nyní udělejme malý pokus – vezměme si tabulku, která obsahuje průměrné frekvence slov ve větách o sedmi až jedenácti slovech (stále korpus CHALDUN):

[53] Vlastně máme jednoduchý vzorec n/2 pro sudé i liché *n*, neboť u lichého *n* stejně musíme zaokrouhlovat nahoru na celé tokeny.

Věty o 7 slovech	Věty o 8 slovech	Věty o 9 slovech	Věty o 10 slovech	Věty o 11 slovech
2313,796	2472,702	2613,26	2760,395	2729,088
4994,123	4750,046	4627,127	4780,654	4464,241
6308,766	6300,585	6276,785	6519,475	6597,159
5650,959	5958,027	6152,409	6296,057	6085,309
5950,323	5513,831	5702,836	5788,941	5894,079
7741,494	5917,417	5605,067	5878,076	5789,036
760,9259	7538,283	5832,088	5589,03	5519,646
	783,0388	7623,991	6059,408	5754,243
		739,7853	7411,464	5792,67
			819,2273	7637,717
				751,8218

Nyní z každého sloupce budeme postupně *vypouštět prostřední token*, dokud ze všech sloupců nedostaneme pouze data pro věty o 7 tokenech:

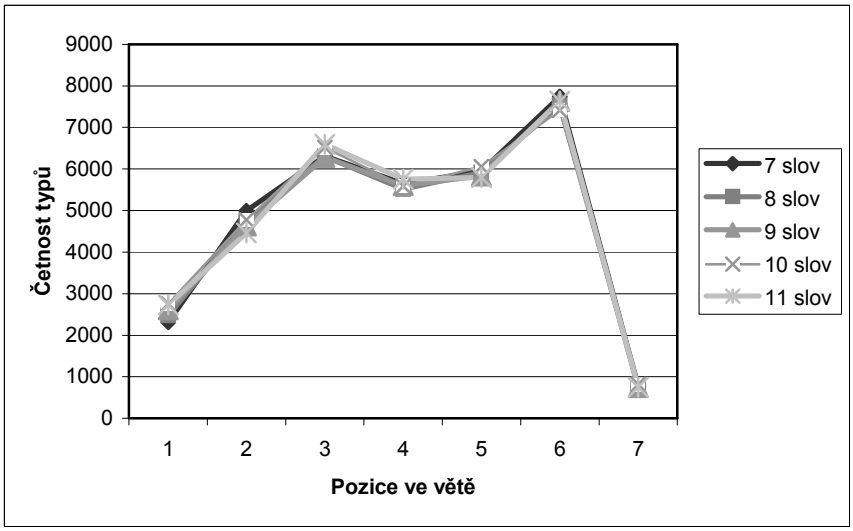
Věty o 7 slovech	Věty o 8 slovech	Věty o 9 slovech	Věty o 10 slovech	Věty o 11 slovech
2313,796	2472,702	2613,26	2760,395	2729,088
4994,123	4750,046	4627,127	4780,654	4464,241
6308,766	6300,585	6276,785	6519,475	6597,159
5650,959				
5950,323	5513,831			
7741,494	5917,417	5605,067		
760,9259	7538,283	5832,088	5589,03	
	783,0388	7623,991	6059,408	5754,243
		739,7853	7411,464	5792,67
			819,2273	7637,717
				751,8218

A tyto datové řady o 7 prvcích „scukneme“ dohromady tak, abychom z nich mohli vytvořit graf[54]:

2313,796	2472,702	2613,26	2760,395	2729,088
4994,123	4750,046	4627,127	4780,654	4464,241
6308,766	6300,585	6276,785	6519,475	6597,159
5650,959	5513,831	5605,067	5589,03	5754,243
5950,323	5917,417	5832,088	6059,408	5792,67
7741,494	7538,283	7623,991	7411,464	7637,717
760,9259	783,0388	739,7853	819,2273	751,8218

A vytvořme ho:

[54] Doufám, že jsem tento postup popsal přehledně, jedná se pořád o stejná data.



Černobílé grafické prostředí nám bohužel nedovoluje znázornit tuto skutečnost esteticky, nicméně myslím, že je dostatečně zřetelně vidět, že se hodnoty prakticky překrývají. Čili že můžeme jednoduchým algoritmem vytvořit z grafu pro n-slovné věty graf pro m-slovné věty, kde $m < n$.

Povzbuzeni tímto vývojem přejdeme k testování hypotézy, která je vlastně preciznějším vyjádřením hypotézy číslo 4:

Hypotéza číslo 5: **Ve frekvenčním zobrazení je průměrná věta o *n* tokenech ekvivalentní průměrné větě o *n* + 1 tokenech s *vypuštěným prostředním tokenem.***

Ověřme si ji napřed na datech získaných opět na korpusu CHALDUN:

Pearsonův korelační koeficient pro data naměřená na větách o 7 tokenech a větách o 8 tokenech s vypuštěným prostředním tokenem činí 0,9989 (dále značíme R (7-8)). Další data budeme přehledně strukturovat do tabulky:

	R (7-8)	R (8-9)	R (9-10)	R (10-11)	R (11-12)
CHALDUN	0,9989	0,9987	0,9978	0,9945	0,9955
ZOLA	0,9967	0,9937	0,9977	0,9969	0,9973
AWU	0,9895	0,9949	0,9807	0,9737	0,9718

Vidíme, že korelační koeficienty jsou velmi vysoké. Podle očekávání se tedy datové řady vzájemně velmi podobají.

Myslím, že s klidným svědomím můžeme uzavřít, že čtvrtá, respektive pátá hypotéza je vyvratitelná a nevyvrácená, ovšem další otázky se jen hrnou, zpracovat je by však zabralo příliš

mnoho času a místa: Dává vypuštění prostředního tokenu vždycky nejlepší výsledky? Neexistuje nějaký sofistikovanější algoritmus, pomocí kterého by korelace byly ještě vyšší? Neliší se tento algoritmus v závislosti na jazyce?

2.5 Shrnutí

V obvyklém způsobu uvažování mnoho lidí chápe pojem *průměrný* ve významu *běžný, typický*. To je dáno tím, že vlastnosti, se kterými se lidé v životě setkávají, se často distribuují podle normálního rozdělení[55]. Je třeba chápat průměr jako číslo, které nám neříká, jak vypadá typický prvek, ale *jak by to vypadalo, kdyby všechny prvky byly stejné*. Strukturám, kterými jsme se v této kapitole zabývali, neodpovídá pravděpodobně ani jedna věta v korpusu, jsou patrné teprve při zprůměrování velkého množství vět a je s tím třeba počítat při vyvozování závěrů (nicméně základní rysy jsou patrné již při zpracování několika desítek vět, jak se ukázalo při práci s verši).

„Velbloudí křivky" můžeme interpretovat tak, že četnost slov je odvozená od slovního druhu a pozice každého slovního druhu ve větě je určena jeho typickou syntaktickou úlohou. Například relativně vzácná slova na počátku arabské věty vysvětlíme tím, že na počátku arabské slovesné věty bývá často sloveso a flektivní slovesa mají mnoho forem, čili se málokdy přesně opakují. Naopak v anglických větách, které typicky začínají zájmenem, kterých je celkově málo, bývá první slovo ve větě nadprůměrně četné.

Tímto způsobem ovšem jen těžko vysvětlíme, proč se méně četná slova tak výrazně soustřeďují na konec věty, a tvoří tak jakési oddělovací značky mezi segmenty. Navíc to může být také přesně naopak, třeba se slovosled přizpůsobuje struktuře, která určuje rytmus, s jakým člověk přirozeně střídá četnější a méně četná slova.

Jisté však je, že vzácná slova na konci významového celku nejsou doménou poezie, že je to obecná vlastnost nejen arabštiny, ale i angličtiny, francouzštiny a češtiny, možná i řady dalších jazyků. Vzhledem k tomu, že data naměřená na větách o určitém počtu slov jsou si velmi podobná s daty naměřenými na větách o slovo delších, můžeme tvrdit, že nejen krajní slova mají pevně stanovenou průměrnou frekvenci, ale že i uprostřed vznikající „velbloudí křivka" není náhodným prvkem, ale že ukazuje ke složitější struktuře, podle které se řídí věty o různých

[55] Například průměrná žena váží 65 kg, a to je skutečně váha, kolik tak normální ženy vážívají. Problém nastává tehdy, když se prvky nedistribuují podle Gaussovy křivky ani podle Poissonova rozdělení. Například peníze ve společnosti sledují, podobně jako slova, mocninnou křivku Zipfova zákona (ovšem ekonomové Zipfovu zákonu říkají Paretův princip), a tak z lidových vrstev často slyšíme nářky, že průměrný plat nikdo nemá a že statistici lžou.

počtech tokenů. Tento obecný princip se potvrzuje také tím, že velmi podobné křivky je možné naměřit na větách různých textů napsaných ve stejném jazyce.

Zvláštní také je, že příbuzné jazyky vytvářejí podobné křivky a že můžeme najít společné rysy u křivek pro všechny zkoumané jazyky. Je otázka, nakolik je to dáno nedostatečným počtem prozkoumaných jazyků a nakolik tu hledíme tváří v tvář obecně lingvistické zákonitosti[56].

Je velmi lákavé aplikovat popsaný algoritmus na další texty a další jazyky, abychom mohli rozhodnout, zdali je tato „velbloudí" povaha křivek obecnou vlastností všech jazyků (musím přiznat, že by mě to příliš nepřekvapilo); avšak záměr a téma práce mě zavazuje odtrhnout se od tohoto tématu a zamířit k dalším úrovním popisu struktury arabského textu.

[56] Je sporné, jestli nám odhalení těchto struktur přineslo nějaký praktický užitek. Na delších textech šifrovaných pomocí kódovacích tabulek pro jednotlivá slova nám sice může pomoci odhadnout znaky pro hranici vět, popřípadě napovědět, jaký je jazyk otevřeného textu, ovšem myšlenka, že v době asymetrických šifer někdo bude kódovat podle tabulky, je poněkud úsměvná. Spíše může napomoci při luštění starých logografických písem. Pokud máme k dispozici do menších celků strukturovaný „text" v neznámém písmu nebo jazyce, pak popsané hypotézy nám mohou napomoci určit, jestli se vůbec jedná o skutečný text v přirozeném jazyce (vede se například diskuze, jestli dosud nerozluštěný Voynichův rukopis není falzum, které se jako skutečný text pouze tváří).

3. Délka slov ve větě

Otázka, co je to vlastně délka tokenu, není tak triviální, jak by se na první pohled zdálo. Musíme mít stále na paměti, že hypotézy, které formulujeme pro psaný text, na němž délku slova měříme v počtu písmen, mohou mít jen volný vztah k zákonům, jež budou platit pro text mluvený, nejen proto, že počet písmen ve slově se často liší od počtu hlásek[57], a nejen proto, že různé hlásky trvají obvykle různou dobu, ale hlavně proto, že samotná délka tokenu závisí na mnoha faktorech, nejen na počtu a jakosti hlásek. Možná, že kdybychom měřili délku tokenu v milisekundách, dospěli bychom ke zcela jiným výsledkům. Z technických důvodů se však prozatím spokojíme s délkou tokenu měřenou počtem písmen, které obsahuje.

3.1 Motivace

Frekvence slova a jeho délka jsou nerozlučně spjaty. To dokládá jednak vztah frekvence slova na jeho délce[58], jednak vztah průměrné frekvence slova ve větě na průměrné délce slova ve větě, což podrobněji řeším v kapitole 4. Nemůžeme si tedy neklást otázku: najdeme podobné struktury, jaké jsme našli v kapitole 2 i v případě, že místo průměrů frekvencí, budeme zkoumat průměry délek jednotlivých slov ve větě? Tato kapitola má všechny předpoklady k tomu, aby byla velmi nudná, neboť v ní přebírám techniky, algoritmy a mírně upravený software z kapitoly předchozí. Budu se ji tedy snažit co nejvíce zkrátit a oživit srovnáváním s výsledky v druhé kapitole.

3.2 Metoda měření

Provedeme **zobrazení textu podle délky slov**: **Každému tokenu v korpusu přiřadíme délku tohoto tokenu.**

---

[57] Pravopis francouzštiny, angličtiny a nevokalizované arabštiny je v tomto ohledu extrémní. Ovšem když jsem zkusil některé algoritmy této kapitoly na německém textu a poté na jeho fonologickém přepisu, výsledky byly prakticky totožné.

[58] Které je zřejmě dáno snahou člověka o jazykovou ekonomii – slova, která je nutno používat často, jsou přirozeně krátká. Dlouhé typy, kterých (z kombinatorických příčin) může být více, naopak obsazujeme významy s řídkým použitím, což se občas odráží i v diachronní perspektivě. Ovšem takové filozofování není účelem této studie. Preciznější formulaci vztahu délka slova – jeho frekvence, najdete v Encyklopedii Laws in Quantitative linguistics v článku Word length and frequency (Strauss – Altmann 2006)

Poté roztřídíme věty podle jejich délky. Dostaneme tak data, jako například v této tabulce[59]:

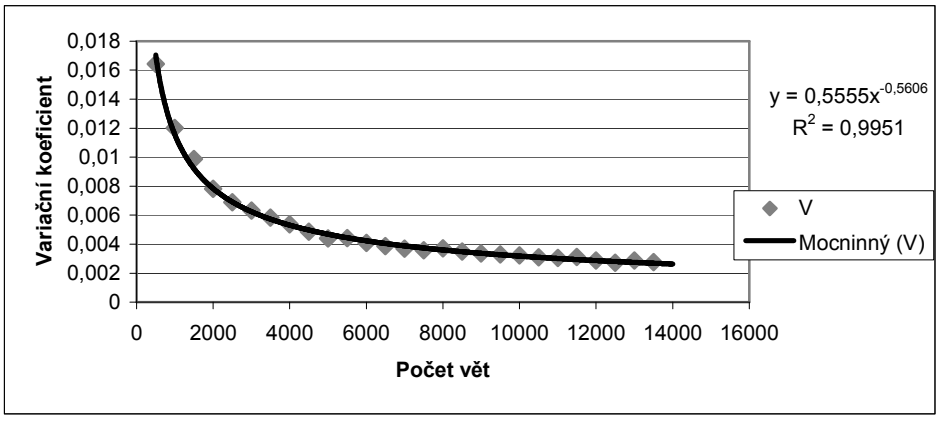
Původní text (zprava doleva)	Délka	Zobrazení (zleva doprava)
ديار لسلمى عافيات بذي خال ألح عليها كل أسحم هطال	10	4 5 6 3 3 3 5 2 4 4
وتحسب سلمى لا تزال ترى طلا من الوحش أو بيضا بميثاء محلال	12	5 4 2 4 3 3 2 5 2 4 6 5
وتحسب سلمى لا نزال كعهدنا بوادي الخزامى أو على رس أوعال	11	5 4 2 4 6 5 7 2 3 2 5
ليالي سلمى إذ تريك منصبا وجيدا كجيد الرئم ليس بمعطال	10	5 4 2 4 5 5 4 5 3 6
ألا زعمت بسبابة اليوم أنني كبرت وأن لا يحسن اللهو أمثالي	11	3 4 6 5 4 4 3 2 4 5 6
كذبت لقد أصبى على المرء عرسه وأمنع عرسي أن يزن بها الخالي	12	4 3 4 3 5 4 5 4 2 3 3 6

Vidíme, že data na rozdíl od frekvence hezky oscilují kolem průměru.

Poté zprůměrujeme hodnoty na jednotlivých pozicích pro věty o stejných počtech slov. Například průměrná řada pro věty o deseti slovech by v této tabulce byla:

4,5 4,5 4 3,5 4 4 4,5 3,5 3,5 5

Abychom mohli porovnávat výsledky s předchozí kapitolou, variační koeficient opět zjistíme experimentálně (Ilustrační data opět pocházejí z korpusu JM s tokeny náhodně rozdělenými do „vět“):

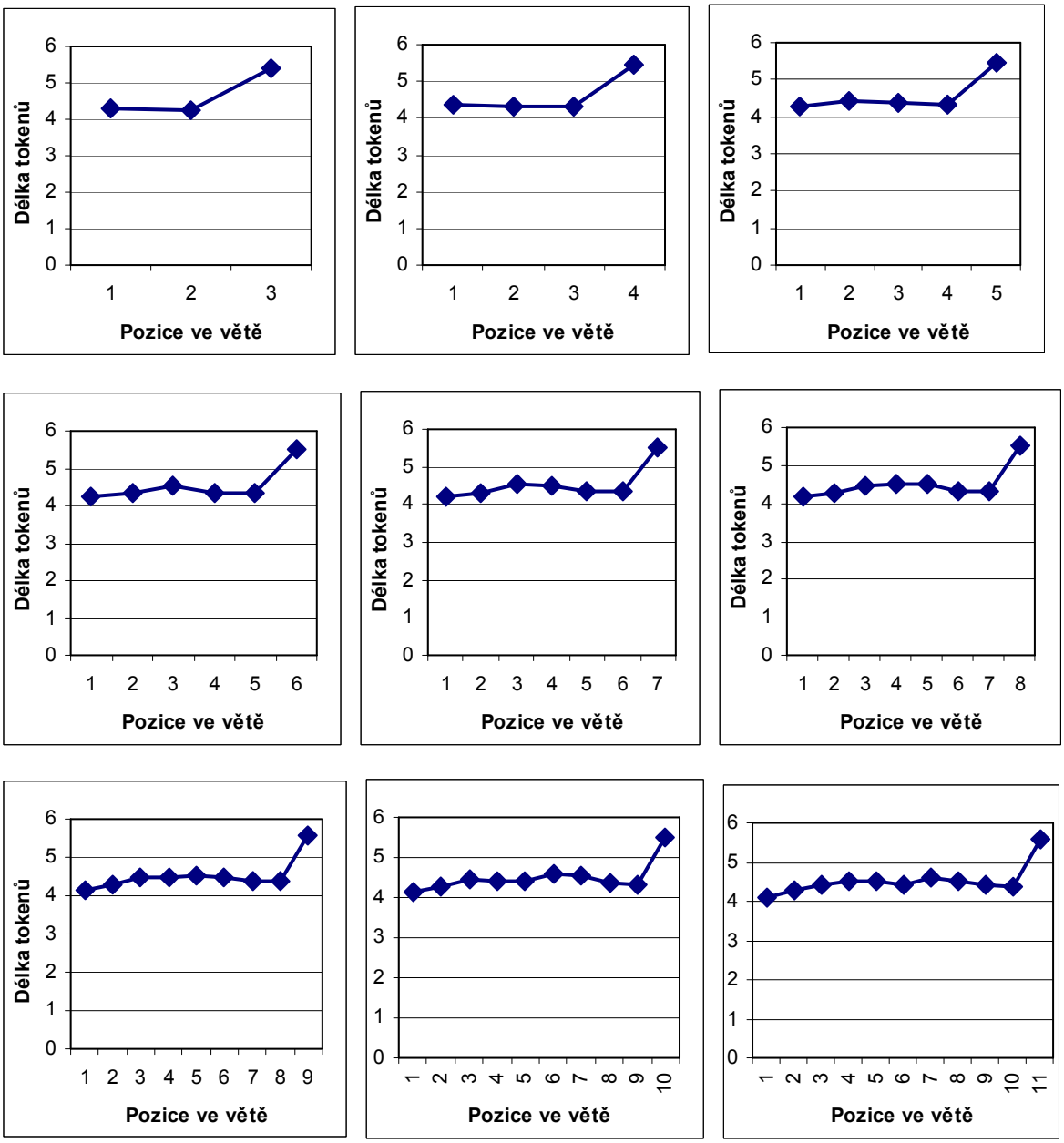


Vidíme, že variační koeficient je o mnoho menší než v prvním případě a že dobře odpovídá křivce mocninného vztahu.

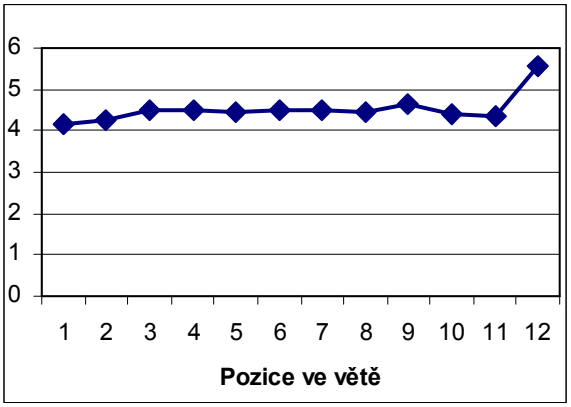
[59] Jedná se o tabulku stejných veršů, jaké jsme užili v kapitole 2.2, aby srovnání bylo dokonalé.

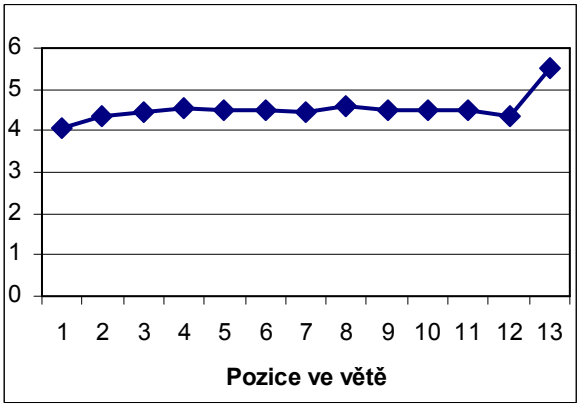
3.3 Aplikace této metody

V krátkosti se podívejme, jak se budou lišit křivky průměrné délky od těch, které jsme získali při měření průměrné frekvence[60].



[60] Stejně jako v předchozí kapitole použijeme korpus AWU, proto jsou počty vět stejné jako v předchozí kapitole: 7300 u prvního grafu (variační koeficient činí přibližně 0,36 %), u druhého 6800 (variační koeficient činí přibližně 0,38) a u třetího 5900 (variační koeficient činí přibližně 0,41), vět o šesti až osmi slovech bylo mezi pěti a třemi tisíci (variační koeficient se pohybuje mezi 0,45 % a 0,6 %), a vět o devíti až jedenácti slovech bylo mezi 2500 až 1500 (variační koeficient se pohybuje mezi 0,7 % a 1 %), vět o 13 slovech bylo 1000 (variační koeficient odpovídá přibližně 1,2 %).





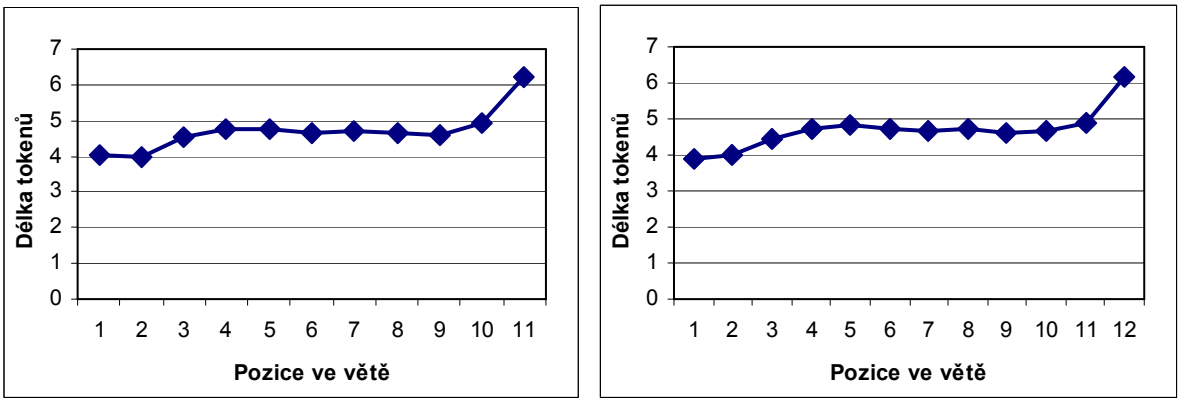
Podle očekávání (nižší frekvence odpovídá delším slovům) je poslední token ve větě delší než ostatní, ovšem na první pohled vidíme, že mechanické převádění křivek pro frekvenci na křivky pro délku nebude možné. Jednak plateau ve střední části je delší a plošší, než jak jsme ho pozorovali u frekvenčního zobrazení, jednak první token je v arabské větě (opticky) průměrně o trochu kratší, než je průměr, ale frekvenci má také menší než průměrnou (viz kapitolu 2.3). „Velbloudí struktury“, na které jsme byli zvyklí, se tak nevytváří. Obdobný „had se vztyčenou hlavou“ se ovšem objeví i u anglického textu[61]:

Lze konstatovat, že věta o 5 tokenech má podobný graf jako u arabského textu, podobnost je opticky větší než u frekvenčního zobrazení. Jak se bude chovat český text? Abychom trochu oživili výklad, nebudeme slepě kopírovat předchozí kapitolu uváděním grafů pro nízké počty tokenů ve větě, neboť malý variační koeficient nám dovoluje dívat se i na věty o větším počtu tokenů, na nichž uvidíme, jak struktura vypadá, když se rozvine[62].

---

[61] Korpus COOPER, v prvním případě 5400 tříslovných vět, pak 5900 pětislovných vět, 5700 devítislovných vět, ve všech případech je variační koeficient okolo 0,74 %.

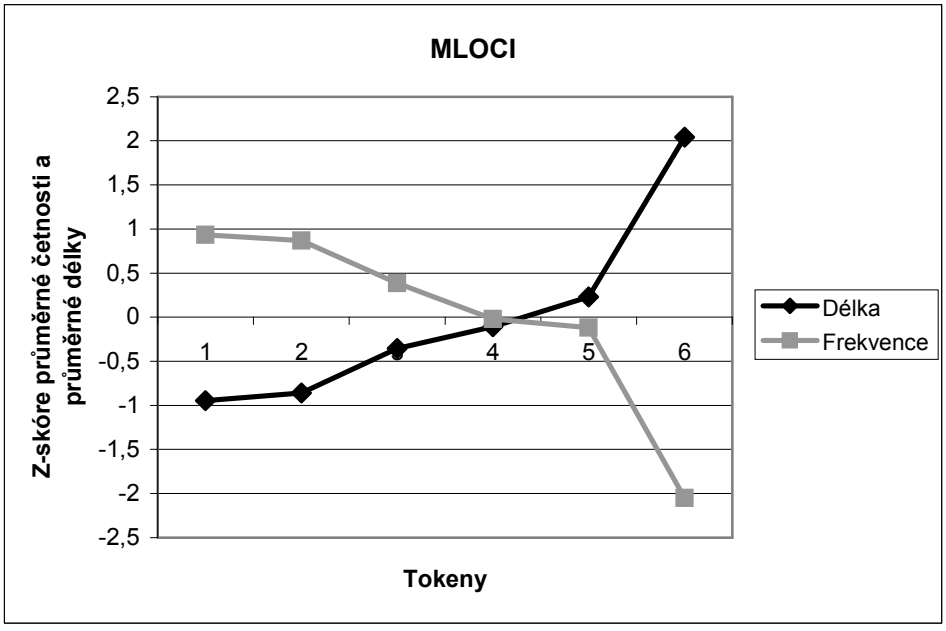
[62] Data byla získána na českém korpusu MLOCI. Data prvního grafu byla změřena na 1700 větách (odpovídá variačnímu koeficientu přibližně 1,25 %), druhého na tisíci větách (odpovídá variačnímu koeficientu přibližně 1,5 %).



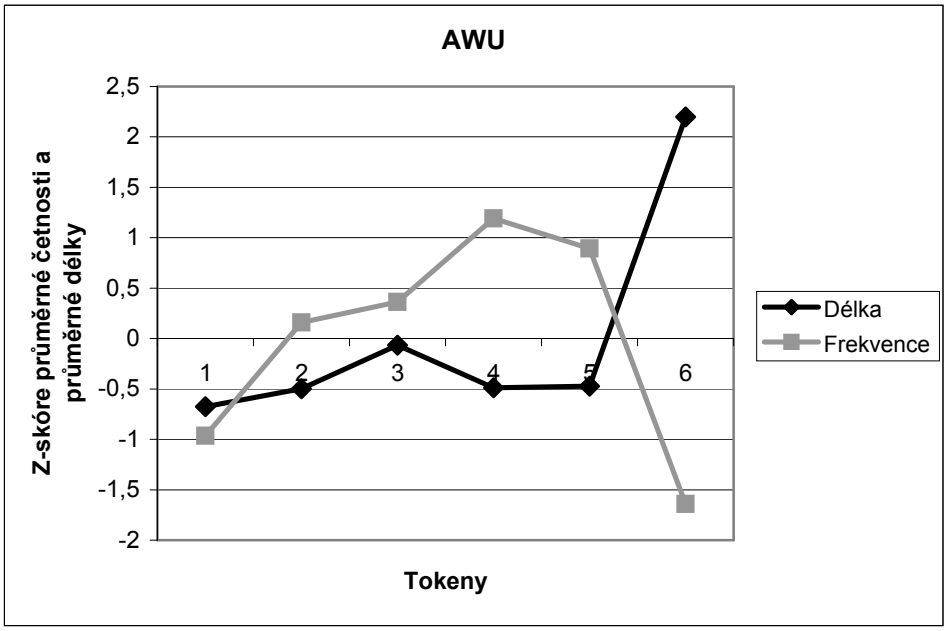
Přestože data byla získána na zcela jiných větách, vidíme, že grafy jsou si vzájemně podobné. Tím se dostáváme k otázce, zda budou platit podobné hypotézy, kterými jsme popsali věty ve frekvenčním zobrazení.

3.4 Interpretace

Nejprve otestujeme hypotézu, kterou jsme naznačili na začátku kapitoly 3.1, když jsme mluvili o tom, že frekvenční a délkové zobrazení spolu souvisí. Necháme se přitom inspirovat následujícími grafy[63]:



[63] Aby byla datová řada pro délkové zobrazení souměřitelná s datovou řadou pro frekvenční zobrazení, obě jsem je normalizoval na Z-skóre, tedy od hodnoty pro každou pozici ve větě jsem odečetl průměr pro všechny pozice a vydělil směrodatnou odchylkou (hodnoty jednotlivých pozic se řídí normálním rozdělením, takže je to povolená operace). Měřeno na šestislovných větách korpusu MLOCI. Data druhého grafu měřena na šestislovných větách korpusu AWU.

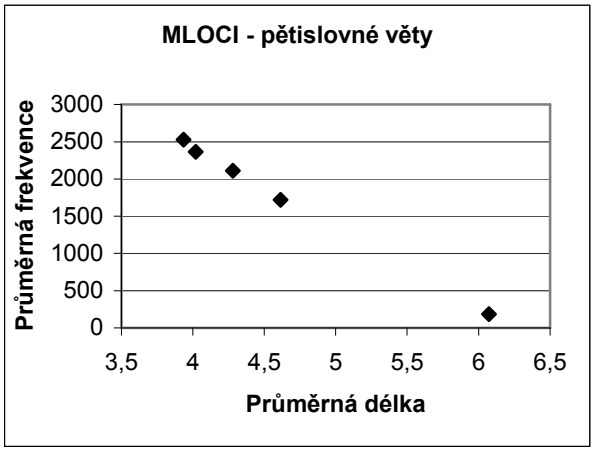


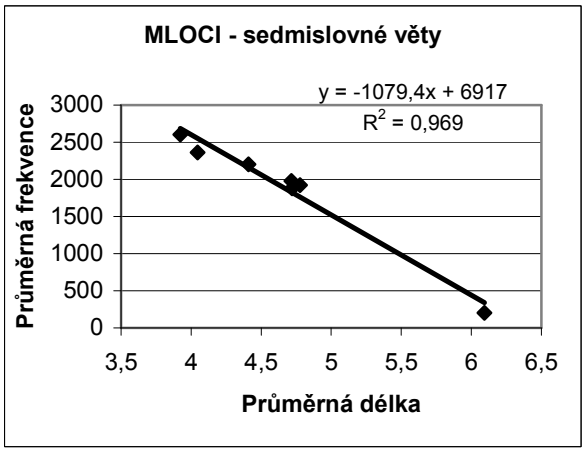
Jak pozorujeme na českém textu, někdy je graf pro délkové zobrazení s jeho ekvivalentem naměřeným na délkovém zobrazení téměř osově souměrný (podle osy *x*), jindy, jako například na arabském korpusu AWU, je vztah složitější (ovšem není to specifikum arabského jazyka, CHALDUN a JM se chovají méně chaoticky). Nemůžeme tedy jednoduše uzavřít naše úvahy s tím, že co platí pro frekvenční zobrazení, bude platit i pro zobrazení délkové.

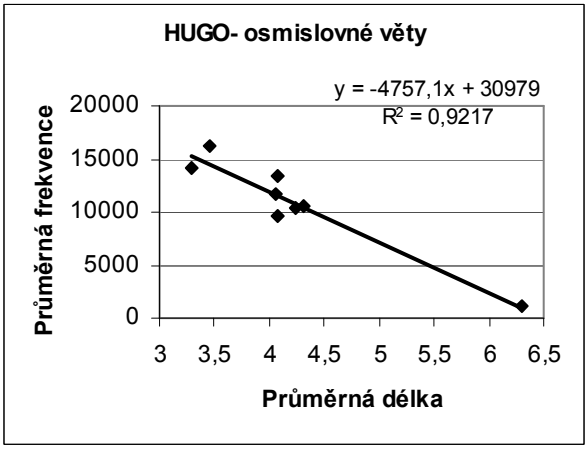
Vzhledem k tomu, že má-li slovo vyšší frekvenci, má průměrně nižší délku, očekáváme mezi „velbloudími" a „hadími" křivkami[64] vysoký záporný korelační koeficient. Pro vztah mezi délkou slova a jeho frekvencí ovšem nepředpokládáme čistou nepřímou úměru (to je také důvod, proč jsme ho formulovali tak opatrně). Můžeme tedy použít Pearsonova koeficientu, který pracuje s lineární korelací? Podívejme se na bodový graf, ve kterém jsou na ose *x* vyneseny průměrné délky slov na jednotlivých pozicích ve větě a na ose *y* jejich průměrné frekvence[65]:

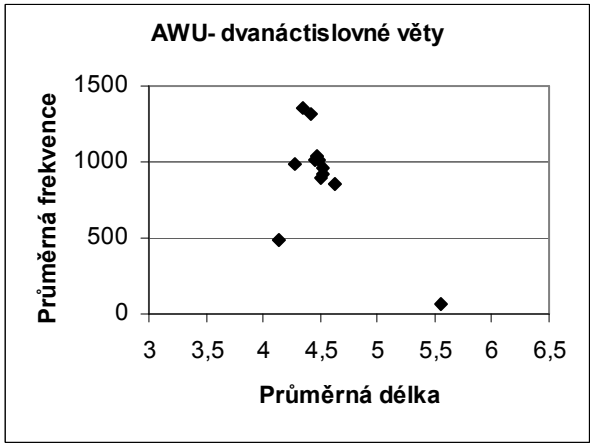
[64] Mezi grafy získanými z průměrných frekvencí slov na jednotlivých pozicích ve větě a grafy získanými z průměrných délek slov na jednotlivých pozicích ve větě.

[65] Použit korpus MLOCI. Každý datový bod značí jeden token ve větě, proto má graf pro pětislovné věty pět bodů, pro sedmislovné sedm bodů. Dále ilustrujeme situaci na osmislovných větách korpusu HUGO a dvanáctislovných větách korpusu AWU.









Vidíme, že vztah někdy sedí lépe, jindy hůře, nicméně v mnoha případech (arabský korpus AWU jsem uvedl jako výjimku) je jeho povaha lineární. Diskuzi o tomto fenoménu přenecháme na konec kapitoly 4.2. Z toho ovšem vyplývá, že můžeme s klidným svědomím použít Pearsonův korelační koeficient pro srovnávání datových řad, jež jsme získali při frekvenčním zobrazení, s těmi, které jsme naměřili stejnou metodou na délkovém zobrazení, a tak formulovat načisto další hypotézu:

Hypotéza číslo 6: **Existuje významná záporná korelace mezi průměrnou frekvencí slova na pozici *m* ve větách o *n* tokenech a průměrnou délkou slova na pozici *m* ve větách o *n* tokenech.**

Tuto hypotézu můžeme podpořit shrnutím měření korelačních koeficientů na dostupných korpusech. Následující tabulka ukazuje Pearsonův korelační koeficient, srovnávající datové řady, které jsme získali při frekvenčním zobrazení, s těmi, které jsme naměřili stejnou metodou na délkovém zobrazení. Začínáme na větách o třech slovech a končíme na čtrnáctislovných větách.

	3 slova	4 slova	5 slov	6 slov	7 slov	8 slov	9 slov	10 slov	11 slov	12 slov	13 slov	14 slov
AWU	-0,808	-0,761	-0,686	-0,677	-0,670	-0,691	-0,642	-0,697	-0,627	-0,662	-0,588	-0,625
COOPER	-0,908	-0,982	-0,939	-0,910	-0,908	-0,888	-0,881	-0,890	-0,856	-0,859	-0,853	-0,853
ZOLA	-0,998	-0,989	-0,965	-0,958	-0,950	-0,943	-0,930	-0,929	-0,928	-0,926	-0,923	-0,907
HUGO	-0,962	-0,988	-0,970	-0,967	-0,959	-0,960	-0,953	-0,941	-0,942	-0,947	-0,930	-0,937
MLOCI	-0,995	-0,999	-0,999	-0,997	-0,984	-0,961	-0,950	-0,919	-0,927	-0,835	-0,847	-0,898
JM	-0,911	-0,839	-0,767	-0,720	-0,726	-0,700	-0,704	-0,654	-0,682	-0,640	-0,667	-0,679
CHALDUN	-0,897	-0,846	-0,823	-0,790	-0,787	-0,759	-0,738	-0,734	-0,722	-0,705	-0,684	-0,673

Korelační koeficient u všech korpusů roste s rostoucím počtem slov ve větě, avšak můžeme shrnout, že kromě jednoho případu nebyl vyšší než –60 %.

V této kapitole dále upravíme hypotézy, které jsme nalezli v kapitole 2.4 na frekvenčním zobrazení tak, aby je bylo možné testovat na délkovém zobrazení. To také následně provedeme, a to na stejných korpusech jako v předchozím případě, aby vynikly rozdíly a shody.

Hypotéza číslo 7: **Poslední token ve významovém celku má větší průměrnou délku, než je průměrná délka tokenu v textu.**

Podívejme se, jak tato hypotéza platí pro různé jazyky[66]:

Korpus	**Poslední token[67]**	**Předposlední token[68]**	**Průměrný token[69]**
AWU	5,49	4,32	4,57
COOPER	5,86	3,96	4,35
MLOCI	6,07	4,55	4,66
NAHODNY	4,26	4,28	4,27
JM	5,12	3,99	4,24
CHALDUN	5,43	4,19	4,54
HUGO	6,27	3,54	4,33
ZOLA	6,35	3,65	4,44

Na jednu stranu poměr mezi posledním a předposledním, respektive průměrným tokenem není tak výrazný jako v případě četností, avšak stále znatelný[70].

---

[66] Popisy jednotlivých korpusů naleznete v kapitole Charakteristika použitých korpusů, korpus NAHODNY vznikl náhodnou transpozicí tokenů v korpusu AWU (zkráceném) a následným náhodným rozdělením do „vět“.

[67] Průměrná délka posledních tokenů ve větách o třech až dvanácti slovech.

[68] Průměrná délka předposledních tokenů ve větách o třech až dvanácti slovech.

[69] Průměrná délka tokenů ve větách o třech až dvanácti slovech.

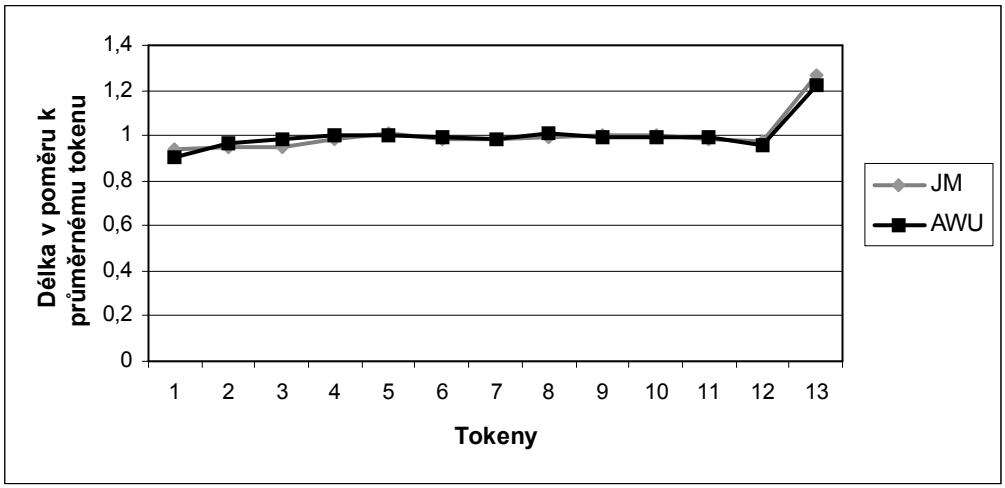
[70] Zkusíme-li použít Studentův test, prozradí nám, že pravděpodobnost, že množina délek posledních tokenů je totožná s množinou délek předposledních tokenů, je prakticky rovna nule (u všech korpusů je menší než rozlišovací schopnost MS Excelu, tedy 10^{-300}, samozřejmě s výjimkou korpusu NAHODNE,

Zkusme si adaptovat další dvě hypotézy z předchozí kapitoly:

Hypotéza číslo 8: **Průměrné věty v textech napsaných stejnými jazyky mají v délkovém zobrazení podobnou strukturu**.

Hypotéza číslo 9: **Průměrné věty v textech napsaných různými jazyky mají v délkovém zobrazení podobnou strukturu.**

Opět se podíváme na tyto dvě hypotézy současně. Napřed uvedeme grafy, které budou ilustrovat devátou hypotézu[71]:

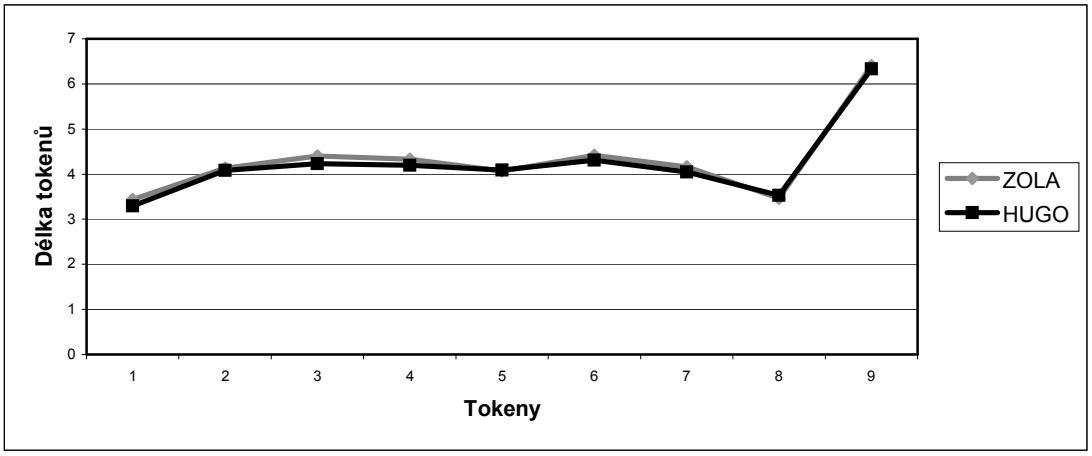


Ještě lépe než v případě frekvence vidíme, že ani tisíc let vývoje jazyka, které tyto dva korpusy dělí, nedokázalo změnit základní struktury, podle kterých se řídí rozložení slov v arabské větě. Křivky se prakticky překrývají. Přesně podle očekávání, ještě lépe dopadlo srovnání korpusů velkých francouzských současníků (ZOLA a HUGO):

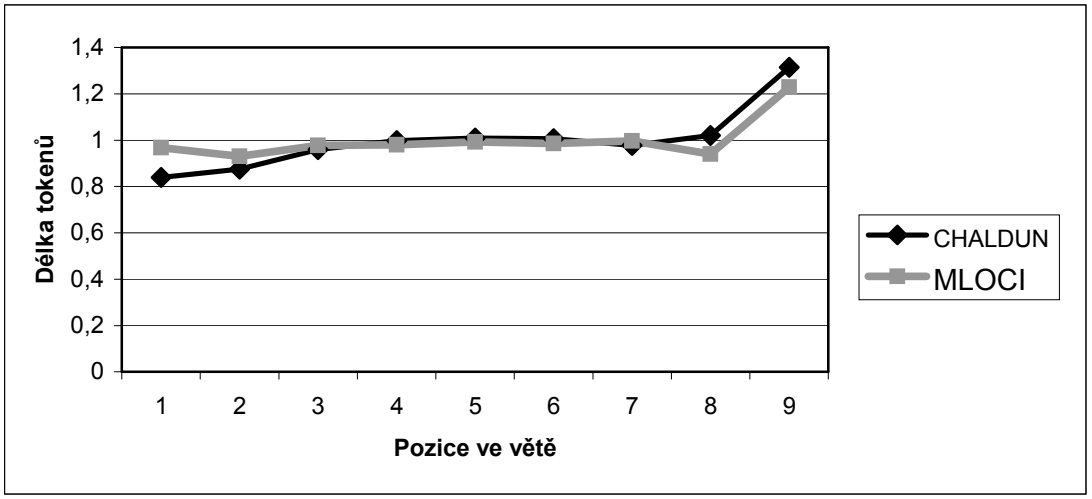
---

kde pravděpodobnost oscilovala mezi deseti a padesáti procenty – test jsem zkoušel několikrát na různých náhodně zpřeházených korpusech). Interpretace tohoto testu je však složitá, neboť délka slov se neřídí přesně podle normálního rozdělení.

[71] Měření bylo provedeno na korpusech JM a AWU, posléze byly výsledky porovnány s průměrnou hodnotou. Korpus JM měl vyšší průměrnou délku slova, zřejmě kvůli občasným vokalizačním značkám, které se někdy do historických textů přidávají, aby modernímu čtenáři usnadnily interpretaci. Dilema – odstranit vokalizaci, a učinit tak korpus lépe porovnatelný s korpusem AWU za cenu ztráty informací, nebo neodstranit a zachovat původní strukturu se vším všudy, jsem rozhodl ve prospěch ponechání vokalizace.



Vidíme, že křivky se překrývají, aniž bychom je museli normalizovat vůči průměru. K ilustraci první hypotézy použijeme opět srovnání korpusu MLOCI a CHALDUN



Navzdory areálním a typologickým rozdílům se křivka pro český korpus podobá křivce pro arabský korpus. Také nyní přejdeme od optického srovnávání k méně subjektivnímu porovnávání korpusů pomocí korelačního a determinačního koeficientu.

Nejprve uvedeme Pearsonův koeficient pro devítislovné věty:

	AWU	**COOPER**	**MLOCI**	**NAHODNY**	**JM**	**CHALDUN**	**HUGO**	**ZOLA**
AWU	**1,0000**	**0,9615**	**0,9618**	*0,3005*	**0,9720**	**0,9609**	**0,9714**	**0,9570**
COOPER		**1,0000**	0,8983	*0,2966*	**0,9134**	0,8889	**0,9655**	**0,9515**
MLOCI			**1,0000**	*0,2396*	**0,9420**	0,8996	0,8936	0,8659
NAHODNY				**1,0000**	*0,3141*	*0,3434*	*0,4559*	*0,4980*
JM					**1,0000**	**0,9827**	**0,9256**	**0,9128**
CHALDUN						**1,0000**	**0,9286**	**0,9248**
HUGO							**1,0000**	**0,9958**
ZOLA								**1,0000**

Dále tatáž data srovnáme pomocí determinačního koeficientu R^2:

	AWU	COOPER	MLOCI	NAHODNY	JM	CHALDUN	HUGO	ZOLA
AWU	**1,0000**	**0,9244**	**0,9250**	*0,0903*	**0,9448**	**0,9232**	**0,9437**	**0,9159**
COOPER		**1,0000**	0,8070	*0,0880*	0,8343	0,7902	**0,9321**	**0,9053**
MLOCI			**1,0000**	*0,0574*	0,8874	0,8093	0,7985	0,7498
NAHODNY				**1,0000**	*0,0987*	*0,0000*	*0,2078*	*0,2480*
JM					**1,0000**	**0,9656**	0,8567	0,8332
CHALDUN						**1,0000**	0,8622	0,8552
HUGO							**1,0000**	**0,9917**
ZOLA								**1,0000**

Opět si v tabulce přehledně uvedeme variační koeficienty pro jednotlivé korpusy:

Korpus	**Počet vět o devíti slovech**	**Variační koeficient (%)**
AWU	2372	0,6
COOPER	4508	0,8
MLOCI	2924	1
NAHODNY	1854	0,8
JM	4966	0,4
CHALDUN	11463	0,3
HUGO	10917	0,3
ZOLA	10900	0,3

Vzájemné srovnávání dvanáctislovných vět bude přesvědčivější než v kapitole 2.4, neboť variační koeficient, v porovnání s devítislovnými větami, nevzroste tak markantně. Tady je Pearsonův korelační koeficient:

	AWU	COOPER	MLOCI	NAHODNY	JM	CHALDUN	HUGO	ZOLA
AWU	**1,0000**	**0,9667**	**0,9146**	*-0,0383*	**0,9623**	**0,9653**	**0,9713**	**0,9739**
COOPER		**1,0000**	0,8690	*-0,0303*	**0,9190**	**0,9062**	**0,9510**	**0,9499**
MLOCI			**1,0000**	*-0,0002*	**0,9303**	0,8810	0,8382	0,8333
NAHODNY				**1,0000**	*-0,0949*	*-0,0587*	*-0,0182*	*-0,0208*
JM					**1,0000**	**0,9717**	**0,9044**	**0,9080**
CHALDUN						**1,0000**	**0,9282**	**0,9324**
HUGO							**1,0000**	**0,9979**
ZOLA								**1,0000**

A zde R^2:

	AWU	COOPER	MLOCI	NAHODNY	JM	CHALDUN	HUGO	ZOLA
AWU	**1,0000**	**0,9345**	**0,8365**	*0,0015*	**0,9260**	**0,9318**	**0,9434**	**0,9484**
COOPER		**1,0000**	0,7551	*0,0009*	**0,8446**	**0,8212**	**0,9043**	**0,9024**
MLOCI			**1,0000**	*0,0000*	**0,8655**	0,7762	0,7025	0,6944
NAHODNY				**1,0000**	*0,0090*	*0,0034*	*0,0003*	*0,0004*
JM					**1,0000**	**0,9441**	**0,8180**	**0,8244**
CHALDUN						**1,0000**	**0,8615**	**0,8693**
HUGO							**1,0000**	**0,9959**
ZOLA								**1,0000**

V tabulce si přehledně uvedeme variační koeficienty pro jednotlivé korpusy:

Korpus	Počet vět o dvanácti slovech	Variační koeficient (%)
AWU	1204	1,1
COOPER	2770	1
MLOCI	1302	1,4
NAHODNY	1091	1,2
JM	2485	0,7
CHALDUN	6554	0,4
HUGO	6206	0,75
ZOLA	4418	0,85

Podívejme se na uvedené tabulky: první, co nás zaujme, je, že korpus NAHODNY má přesvědčivě nízký determinační koeficient k ostatním korpusům, zatímco koreleační koeficienty vzájemně mezi jednotlivými jazyky stouply (v porovnání s frekvenčním zobrazením), a to nejen u korpusů psaných stejným jazykem (HUGO – ZOLA ve všech případech nad 99 %, pro dvanáctislovné věty je R = 99,8 %; velmi vysoká korelace mezi JM – AWU a JM – CHALDUN), ale dokonce i u korpusů napsaných vzájemně vzdálenými jazyky (AWU – COOPER 96 %, JM – MLOCI 93-94 %). Opět jsme v pokušení naši zkušenost zobecnit a říci, že tyto struktury, jak je pozorujeme, jsou společné všem jazykům, ale opět to neuděláme, neboť jsme jich testovali jen velmi málo.

Spíše můžeme opatrně uzavřít srovnání s tím, že pomocí frekvenčního zobrazení bylo možno odlišit od sebe jednotlivé jazyky, délkové zobrazení dokáže dobře odlišit strukturu přirozeného jazyka od „textu“ této struktury zbaveného.

Pojďme se podívat na další hypotézu podobnou té z předchozí kapitoly:

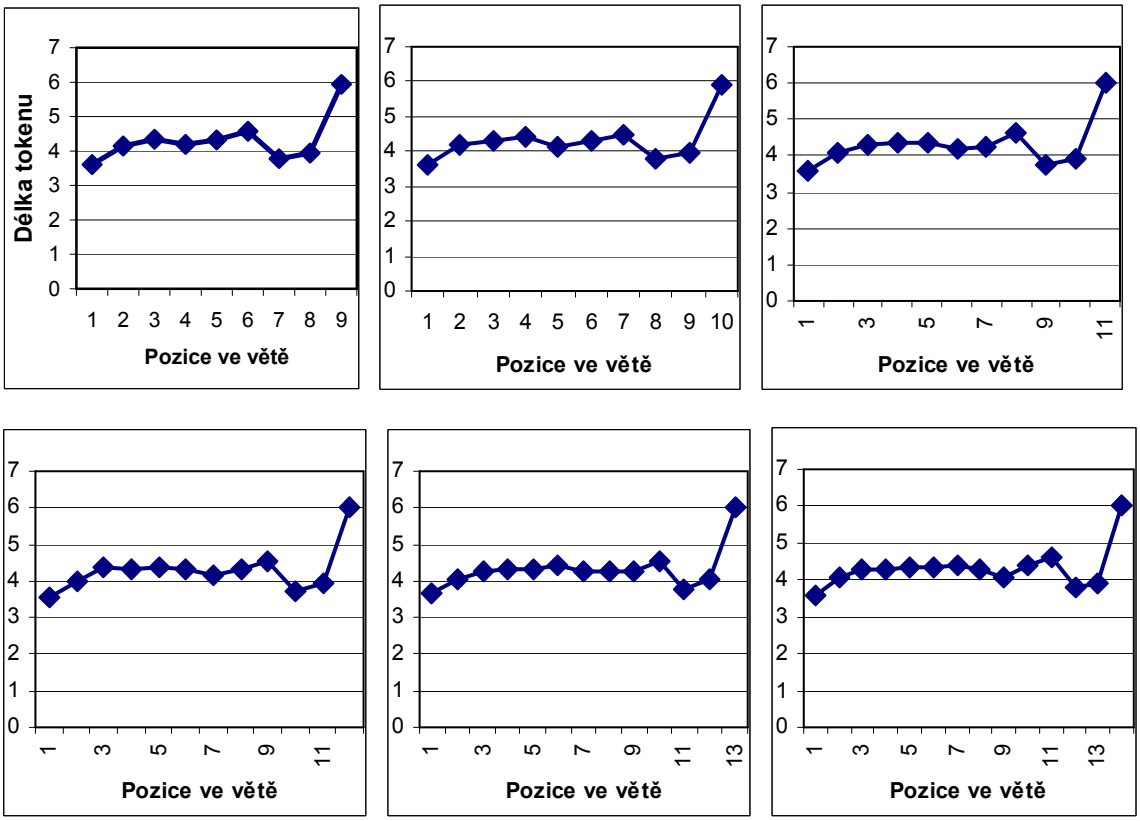
Hypotéza číslo 10: **V rámci jednoho textu mají průměrné věty o různé délce v délkovém zobrazení podobnou strukturu.**

V kapitole jsme si definovali **vypuštění prostředního tokenu** a díky této definici můžeme tuto hypotézu formulovat preciznějí:

Hypotéza číslo 11: **V délkovém zobrazení je průměrná věta o *n* tokenech ekvivalentní průměrné větě o *n* + 1 tokenech s *vypuštěným prostředním tokenem.***

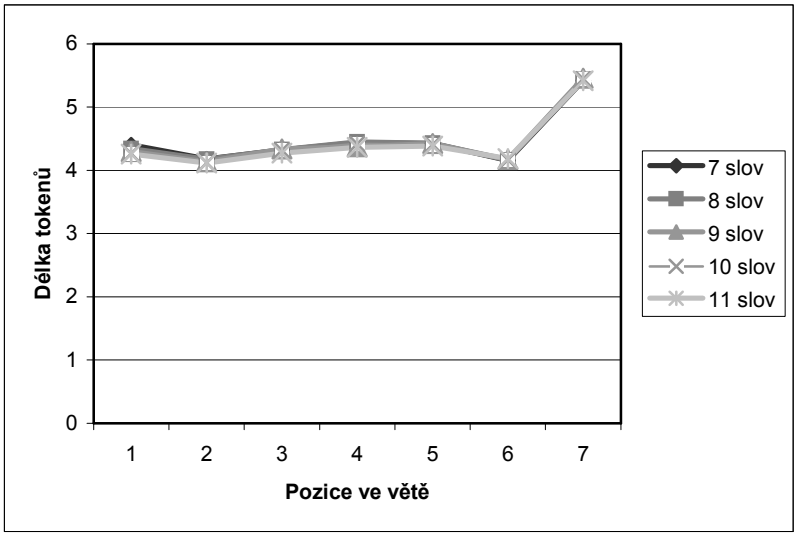
V kapitole 2.4 jsme tyto dvě hypotézy ilustrovali grafy získanými na korpusu CHALDUN. Pro délkové zobrazení se však grafy pro tento korpus téměř neliší od těch, které jsme v kapitole

3.3 naměřili pro korpus AWU. Pro zajímavost si tedy uveďme data pro korpus COOPER, který je poněkud méně fádní[72]:

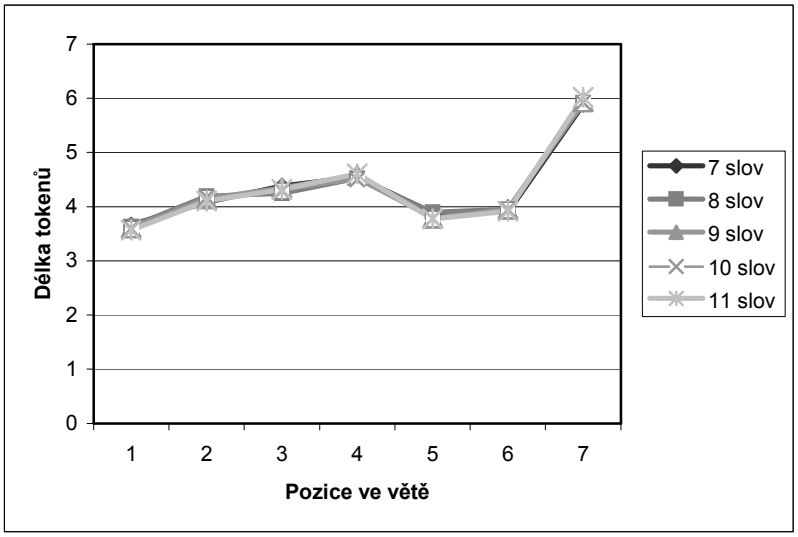


Myslím, že tento sled grafů nás opravňuje k tomu, abychom hypotéze číslo 10 důvěřovali a začali ji testovat. Podobně jako v případě frekvenčního zobrazení provedeme v korpusu CHALDUN ve větách o osmi až jedenácti tokenech tolikrát operaci vypuštění prostředního tokenu, že ze všech těchto vět dostaneme věty o 7 tokenech. Pro bližší vysvětlení odkazuji na kapitolu 2.4, abychom se nezdržovali jeho opětovným vysvětlováním. Pouze uvedeme výsledný graf:

[72] Můžeme si dovolit srovnávat grafy od devítislovných vět výše, neboť díky nízkému variačnímu koeficientu je vliv náhodných výkyvů nízký. Devítislovných vět je 4500, čtrnáctislovných je cca 1800, variační koeficient se pro uvedené grafy tedy pohybuje mezi 0,8 % a 1,3 %.



Opět vidíme, že se řady překrývají. Ještě si ukážeme, jak přesvědčivě bude tento postup fungovat na korpusu COOPER:



Ani tentokrát se nespolehneme na grafy. Pearsonův korelační koeficient pro data naměřená na větách o 7 tokenech a větách o 8 tokenech s vypuštěným prostředním tokenem činí v korpusu COOPER 99,45 % (dále značíme R(7-8)). Další data budeme přehledně strukturovat do tabulky:

	R (7-8)	R (8-9)	R (9-10)	R (10-11)	R (11-12)
CHALDUN	0,9983	0,9972	0,9982	0,9972	0,9927
ZOLA	0,9983	0,9984	0,9992	0,9973	0,9950
AWU	0,9981	0,9972	0,9943	0,9906	0,9795
COOPER	0,9945	0,9930	0,9931	0,9926	0,9970

Korelační koeficienty jsou i tady velmi vysoké. Podle očekávání se tedy datové řady vzájemně velmi podobají, a to většinou více, ale někdy i méně než v případě frekvenčního zobrazení (doporučuji čtenáři výsledky porovnat s těmi na konci kapitoly 2.4).

3.5 Shrnutí

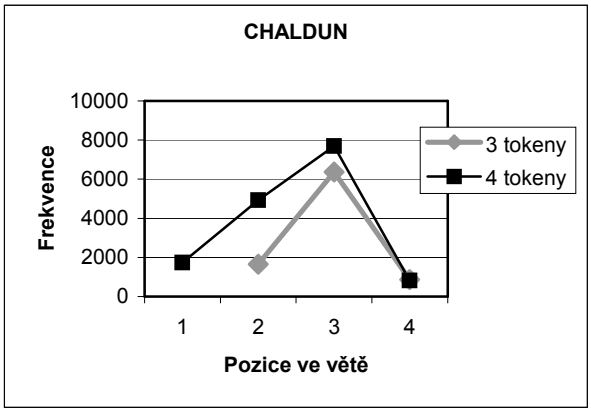
Ukázali jsme si podobně jako v kapitole 2, že délka slov není v textu náhodná, že souvisí s hranicemi významových celků a že v jazyce vytváří jakýsi přirozený rytmus. Tyto rytmy se vzájemně podobají na všech zkoumaných textech, nejvíce na těch, které jsou psány stejným jazykem. Bylo by možná zajímavé pokusit se vyprodukovat text, který by se podle těchto struktur neřídil, tedy pokud by to vůbec bylo technicky možné. Zjistili bychom, jakým způsobem by na něj reagovali příjemci tohoto textu, a jestli by poznali, že je něco v nepořádku, čili jestli tyto struktury hrají v komunikaci nějakou úlohu.

Co je však možná nejdůležitější, tato kapitola názorně ukazuje, že v jazyce jsou pojmy délka a frekvence natolik propojené, že pokud jeden z nich figuruje v nějakém zákoně, či vytváří nějakou strukturu, můžeme očekávat, že podobný zákon nebo podobnou strukturu najdeme i pro druhý člen této dvojice. Také o tom budou pojednávat následující kapitoly.

4. Menzerathův-Altmannův vztah a frekvence slov

4.1 Motivace

Tato kapitola původně vznikala jako poznámka pod čarou u jednoho z grafů v kapitole 2, avšak navzdory očekávání se rozrostla do nynější podoby. Při pohledu na „velbloudí křivky" si při podrobnějším studiu totiž všimneme, že v arabštině má graf pro větu s méně tokeny menší průměrnou frekvenci slov než graf pro větu s více tokeny[73]. Jako například na tomto grafu:



Při pohledu na podobné grafy si vyvodíme jednoduchou tezi:

Průměrná četnost slov ve větě stoupá s její délkou.

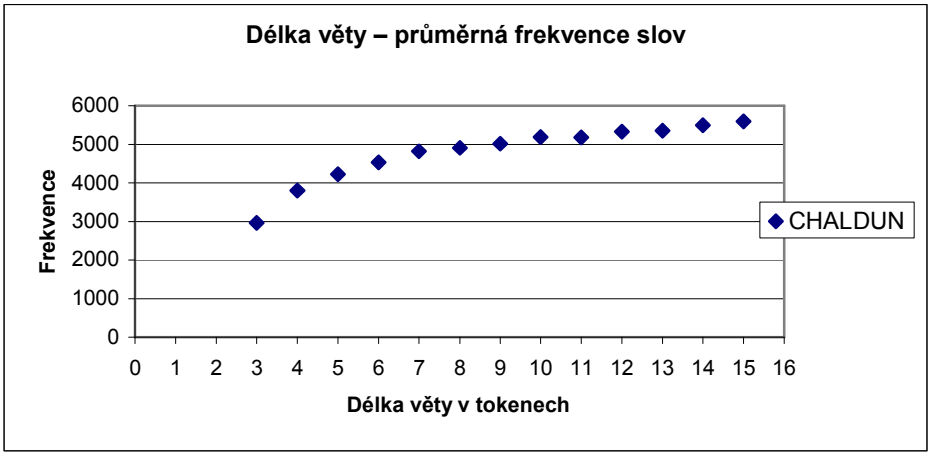
Tuto závislost nebudeme zatím blíže specifikovat. Myslím, že si zaslouží naši pozornost, neboť formuluje vztah mezi různými úrovněmi v jazyce (věta – slovo v ní obsažené), takže nám napovídá, že jazyk se neskládá „z relativně nezávislých subsystémů", ale že naopak jednotlivé vrstvy jsou velmi úzce propojeny.

4.2 Měření a interpretace

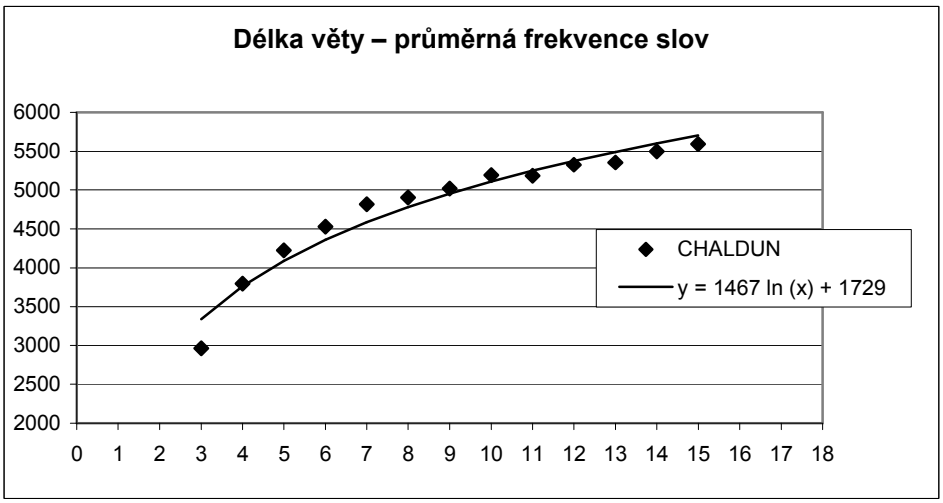
Samozřejmě nedáme na první dojem a podíváme se, jak tento vztah skutečně vypadá. Z předchozích měření máme k dispozici věty roztříděné podle délky (vyjádřeno v tokenech) a

[73] „Kratší velbloud se vleze do delšího velblouda."

k nim přiřazenou frekvenci slov, které obsahují. Stačí nám tedy zprůměrovat tyto frekvence pro celé věty a dostaneme následující bodový graf[74]:

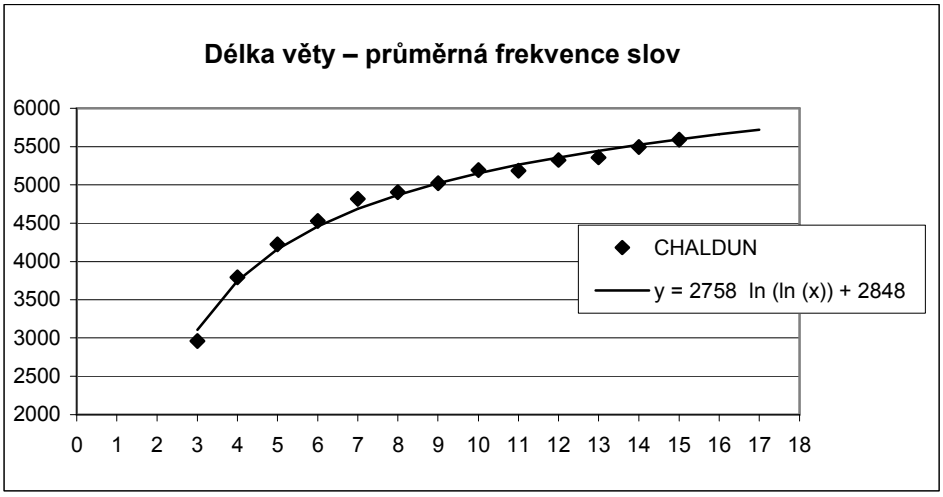


Vyjde nám jedna z oněch křivek, s nimiž se setkáváme v kvantitativní lingvistice velmi často – vypadá to jako mocninný vztah, ale není to mocninný vztah. Pokud bychom se ji pokoušeli modelovat pomocí jednoduchých funkcí, asi nejlépe by mu vyhovovala logaritmická křivka:

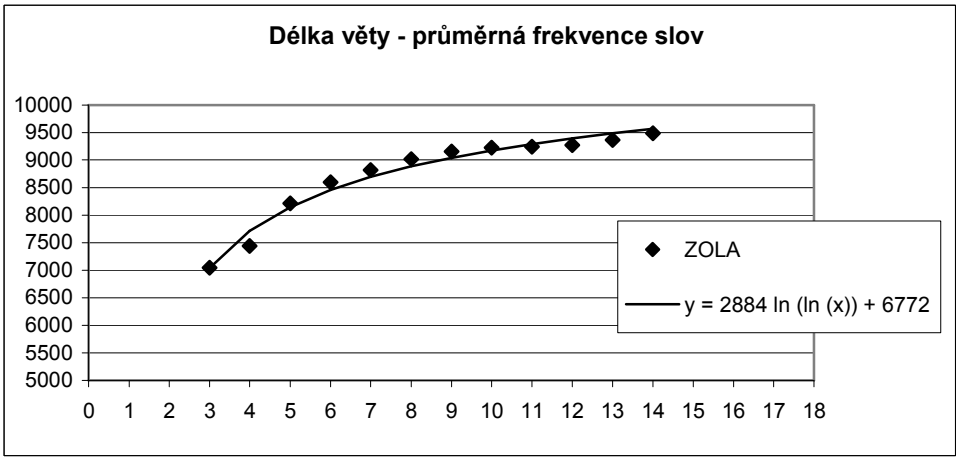


Determinační koeficient R^2 vychází na 0,9531, což není málo, nicméně při pohledu na graf nás tento jednoduchý model neuspokojí. Daleko přijatelnější se opticky zdá, když *x* zlogaritmujeme dvakrát:

[74] Data získána na korpusu CHALDUN.



Také R^2 vychází lépe, na 0,9899, tedy více než v prvním případě. Parametr, kterým dvakrát logaritmované *x* násobíme (2758) má podobnou hodnotu jako parametr, který k rovnici přičítáme (2848), ovšem to je jen náhoda, jak se přesvědčíme u druhého grafu, který jsme získali na datech z korpusu ZOLA[75]:



Zaujme nás, že parametr, kterým logaritmované *x* násobíme, je velmi podobný parametru, který jsme nalezli u Chaldūna. Také toto je náhoda, jiné korpusy mají parametry odlišné.

Obecně si tento vztah vyjádříme jako y = k ln (ln (x)) + a ;

kde *y* je průměrná frekvence slov ve všech větách o *x* slovech, *k* a *a* jsou parametry, které ovšem nejsou charakteristické pro daný jazyk a těžko říci, z čeho vlastně ve skutečnosti vycházejí.

[75] $R^2 = 0{,}9724$

Parametr *a* je o něco menší než průměrná frekvence slov ve větě o třech slovech. To není žádná záhada, když si uvědomíme, že ln (ln (e))=0, tedy vliv parametru *k* je okolo třetího slova minimální[76]. Platí tedy přibližně **a = y (3).**

Najít přibližnou hodnotu parametru *k* také není nic složitého, stačí, když si uvědomíme, že ln (ln (15)) = 0,996 (přibližně 1, neboť e^e = 15,15), takže přibližně platí **a + k = y (15)** – součet obou parametrů je roven průměrné frekvenci slov ve větách o 15 slovech. Velmi zhruba můžeme říci, že y(15) se blíží průměrné frekvenci všech slov v textu (tento výrok jsem ovšem nijak extenzivně netestoval).

Dostáváme se k rovnici (Hypotéza číslo 12):

$y(x)=[y(e^e)-y(e)]\ln(\ln(x)+y(e)$,

která bude mít ve světě přirozených čísel (tokeny nedělíme) poněkud nepřesný tvar:

$y(x)=[y(15)-y(3)]\ln(\ln(x)+y(3)$,

kde za y (15) si můžeme (sice ne s čistým svědomím) dosadit průměrnou frekvenci typu v textu a za y (3) průměrnou frekvenci typů ve větách o 3 slovech. Tento postup hledání parametrů doporučuji pouze ve chvíli, kdy celou řadu neznáme a chceme ji modelovat, aniž bychom ji měřili. Samozřejmě ideální způsob, jak se dostat k co nejpřesnějším parametrům tohoto modelu, je změřit všechny hodnoty na textu a nalézt parametry po dvojitém zlogaritmování osy x a následném dosazení lineárního vztahu tak, aby bylo R^2 co největší.

Nutno dodat, že rovnice pro tento vztah není definitivní, neboť se kdykoli může najít lepší aproximace. To je také důvod, proč kvantitativní lingvista nikdy nepocítí úlevu, jakou nalézali fyzikové, když zformulovali nějaký vztah a mohli si být jistí, že když jim platí doma v pokoji, bude platit i venku na ulici. Tato rovnice také není zrovna intuitivně uchopitelná a těžko říci, jestli je vůbec nějak využitelná. Z praktického hlediska zřejmě nikoli, ale můžeme se snad díky ní hlouběji zamyslet nad vztahem mezi různými úrovněmi jazyka, neboť tento vztah není ve světě osamocen – průměrná délka slova ve větě také závisí na její délce. Tento vztah je znám jako *Menzerathův* zákon, pojďme se na něj podívat blíže, než zjistíme, jestli nebudeme moci nalézt něco, co oba vztahy spojuje.

Menzerathův zákon tvrdí, že: *Čím delší je v jazyce nějaký konstrukt, tím kratší jsou v průměru jeho konstituenty.* (Hřebíček 2002: Str. 53)

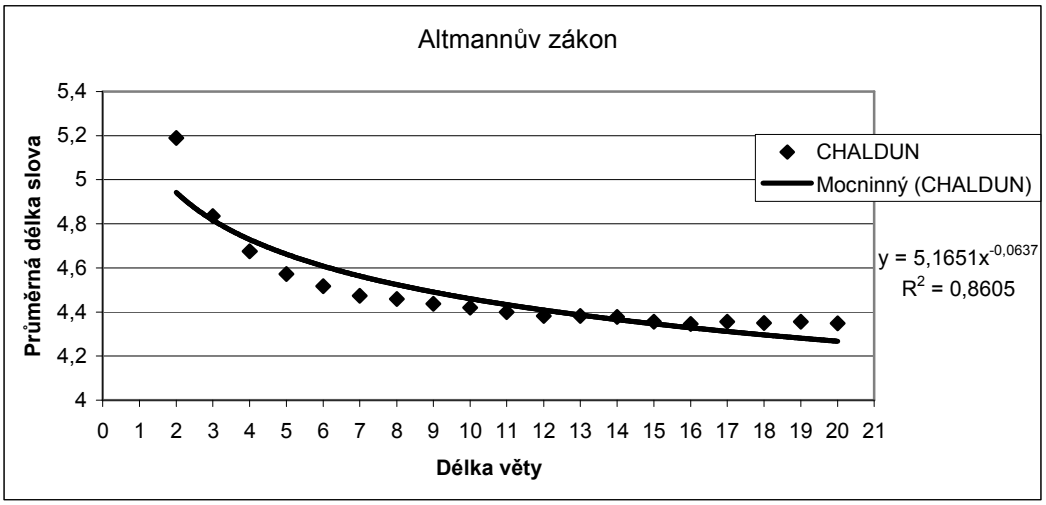
---

[76] Neboť *e* je rovno přibližně 2,7183, takže ln (ln (3)) = 0,094.

Což se podle Gabriela Altmanna dá přesněji vyjádřit vzorcem

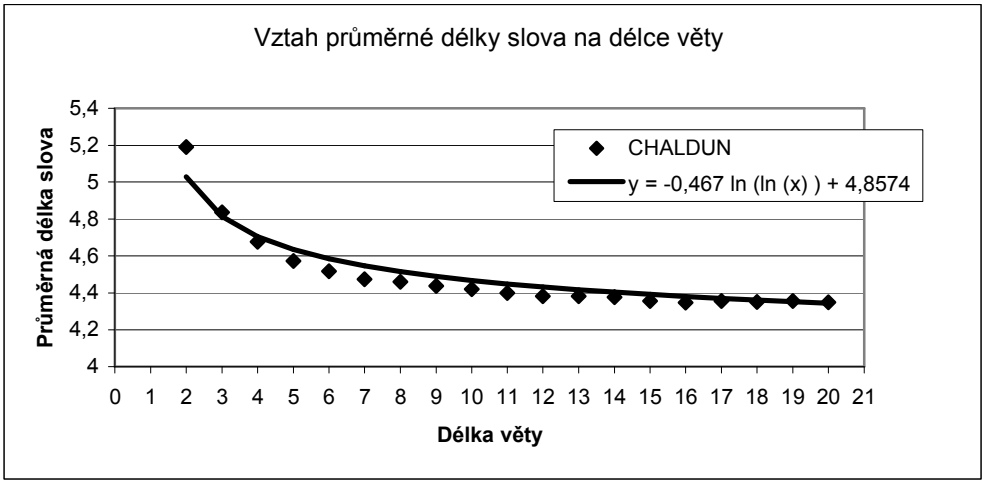
$y = A\ x^{-b}$,

kde *y* je délka konstituentu, *x* je délka konstruktu, *A* a *B* jsou parametry. Takovým konstituentem mohou být například tokeny a konstruktem věty (klauze), jak jsme si je definovali. Je poněkud odvážné formulovat tento vztah jako mocninný, což vidíme na tomto grafu:



Altmann formuloval tento vztah v době, kdy testování kvantitativně lingvistických hypotéz vyžadovalo značnou dávku trpělivosti a vlastně i odvahy. Pro korpus CHALDUN činí $R^2 = 0{,}86$, tedy poměrně nízká hodnota, nicméně tento vztah je v lingvistice uznávaný (nutno říci, že byl původně odvozen pro nižší celky než pro věty a slova a obdivuhodná není ani tak jeho přesnost, jako spíše univerzálnost, s níž popisuje vztah konstituentů a konstruktů na všech úrovních jazyka).

Všimněme si, jak jsou si křivky vztahu, o kterém byla řeč na počátku této kapitoly, a Altmannova zákona podobné. Zkusme nyní uplatnit na vztah délky vět a délky slov obdobný vzorec jako v prvním případě:



Vidíme, že tento vzorec vyhovuje trochu lépe dané křivce a i když výsledek ani nyní nedělá opticky příliš velký dojem, determinační koeficient se zvýšil ($R^2 = 0{,}96$).

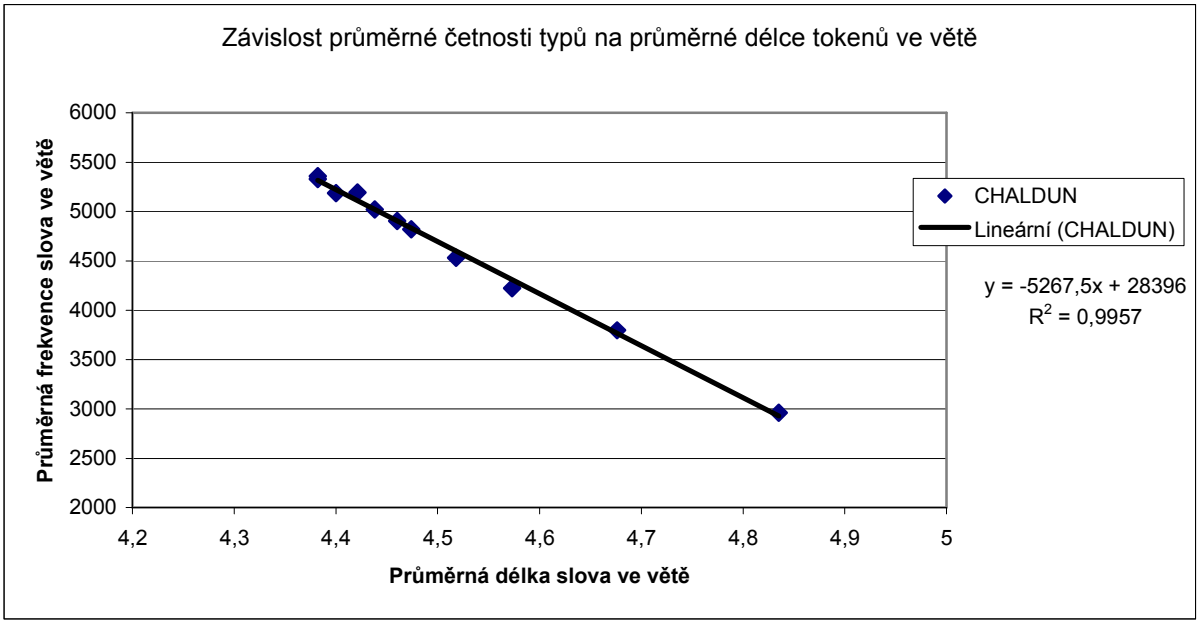
Obecně si tento vztah také vyjádříme jako y = k ln (ln (x)) + a.

Analogicky k prvnímu případu je *y* rovno průměrné délce slova ve větách o *x* slovech, parametr *k* je přibližně roven průměrné délce slova ve větách o třech slovech a *a* + *k* je přibližně rovno průměrné délce slova ve větách o 15 slovech. Samozřejmě je přesnější dospět k hodnotám těchto argumentů statistickými metodami, ovšem vždy je příjemné, když si pod parametry můžeme představit nějaké reálné hodnoty, než když to jsou jen taková neurčitá čísla, ke kterým dospějeme teprve, až když máme vztah změřený.

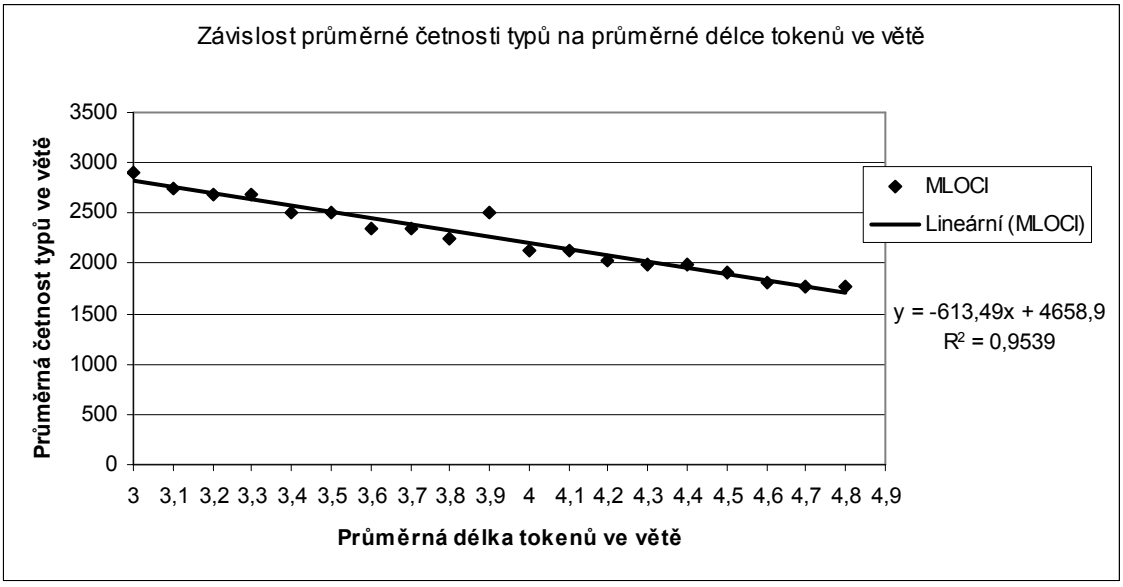
Pokud ovšem tyto dva vztahy (vztah průměrné frekvence slova na délce věty a vztah průměrné délky slova na délce věty) mají stejnou povahu, znamenalo by to, že mezi průměrnou délkou slova ve větě a průměrnou frekvencí slova ve větě je lineární vztah!

Je tomu skutečně tak? Následující graf nám tuto situaci ilustruje[77]:

[77] Do bodového grafu jsem seřadil naměřené výsledky ze závislosti průměrné délky slova na délce věty (osa x) a vztahu závislosti průměrné frekvence slova na délce věty (osa y). Čistě závislost délky typu na frekvenci (tedy nezávisle na členění do vět) popisuje již jednou zmiňované heslo v encyklopedii Laws in Quantitative linguistics (Strauss – Altmann 2006)

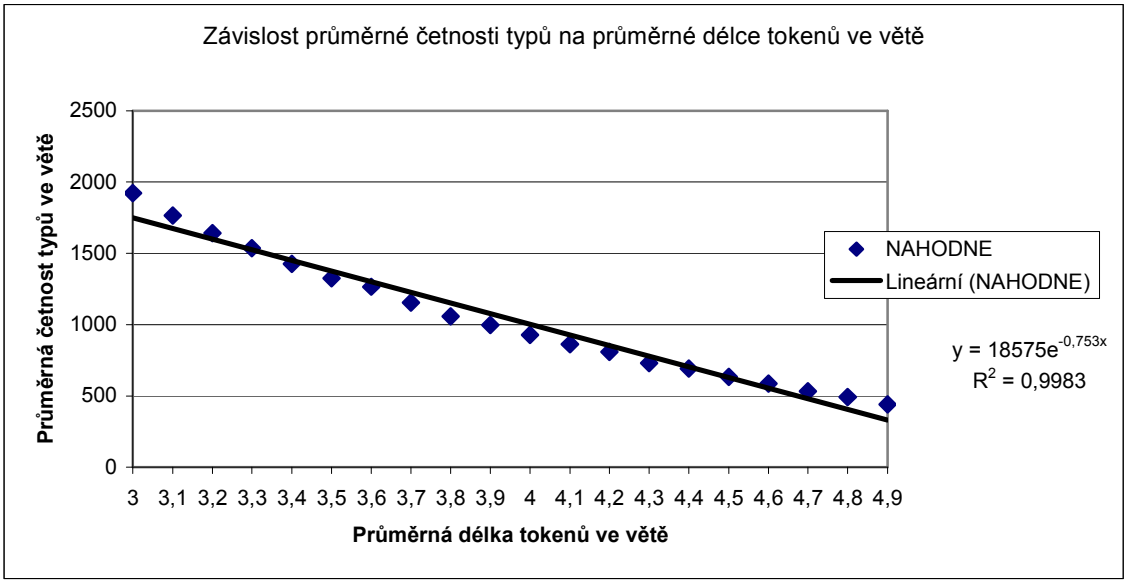


Málokdy se v lingvistice setkáme s tím, že by nějaký vztah šel popsat jako téměř dokonalá lineární závislost. Zkusme si tento vztah ověřit ještě na jiném korpusu, tentokrát změřme celou datovou řadu přímo na textu[78]:

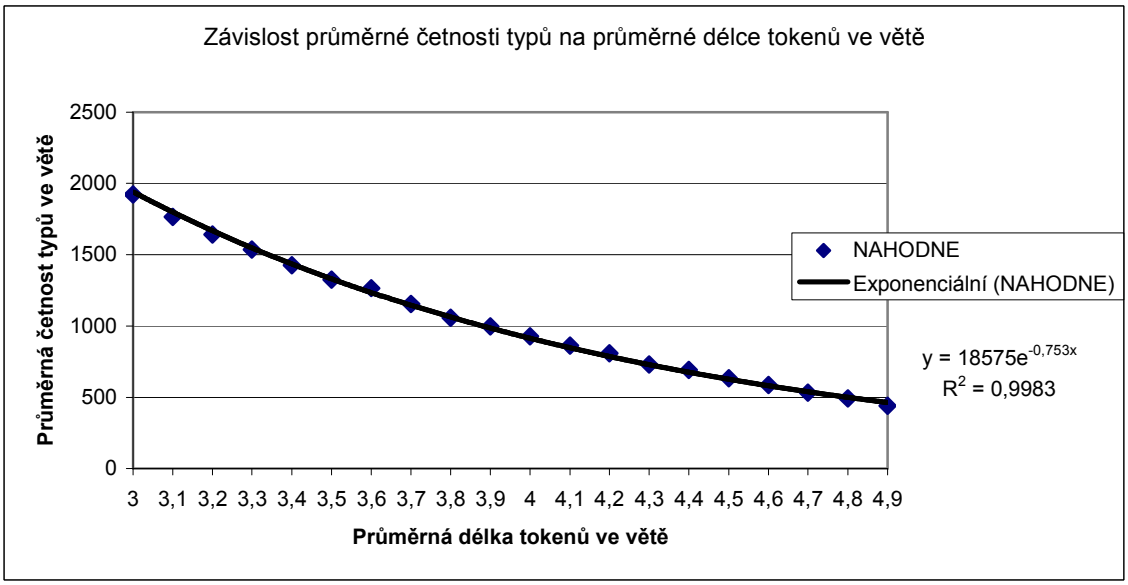


[78] Změřeno na korpusu MLOCI. Ostatní korpusy, které mám k dispozici, také poslušně vytvořily očekávaný lineární vztah, pouze na korpusu ZOLA se mi tento vztah nepodařilo najít, data náhodně oscilovala a nejevila známky jakéhokoli uspořádání. Je velkou záhadou, proč tomu tak bylo, neboť algoritmus byl stejný jako v případě korpusu CHALDUN, HUGO a MLOCI a jinak se korpus ZOLA choval vždy podobně, jako korpus HUGO. Strávil jsem nad tím spoustu času, pokoušeje se osvětlit původ problému, avšak snaha vyšla naprázdno.

Otázkou je, proč se náhodný text chová jinak – na korpusu NAHODNE[79] měla naměřená křivka exponenciální, nebo ještě spíše polynomický charakter. Nejprve se podívejme, jak situace vypadá, když grafem proložíme přímku:

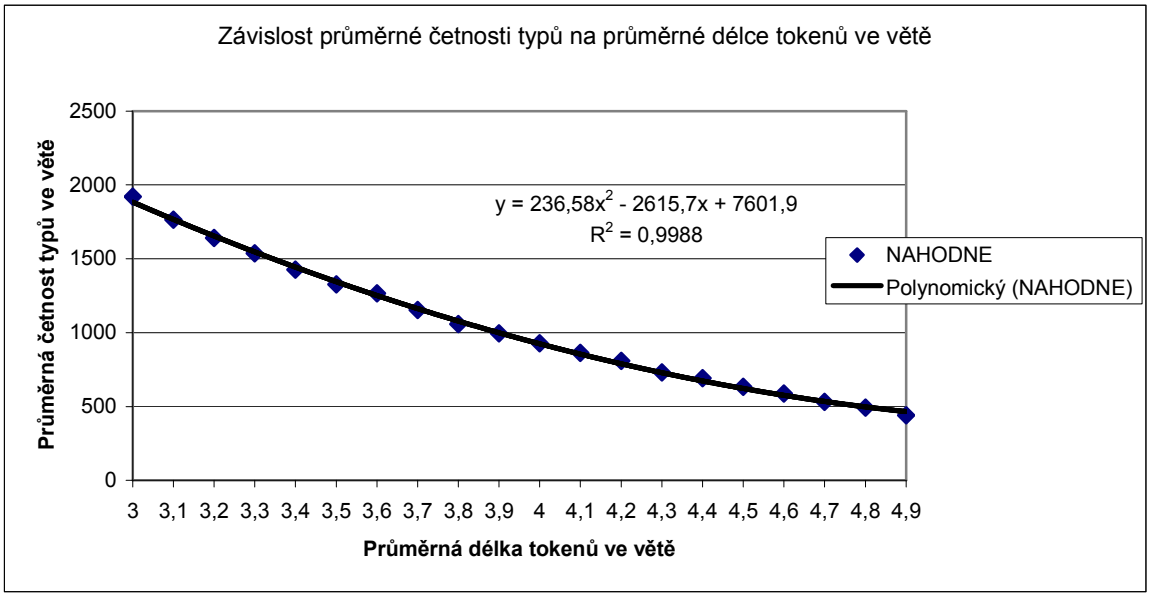


Jak vidíme, determinační koeficient je vysoký, i když křivkou proložíme lineární vztah, který však na první pohled není příliš vhodný. Proložíme-li si křivkou exponenciální funkci, vyjde nám R^2 o něco vyšší:



A poněkud neočekávaně ještě lépe na graf sedí kvadratická funkce:

[79] Korpus JM s náhodně transponovanými tokeny, rozdělený do „vět“ o deseti tokenech.



Otázkou je, jestli bychom téhož nedosáhli při přesnějším měření i na přirozených textech o větším rozsahu, nicméně pro naši potřebu není nutné opouštět myšlenku, že vztah je lineární, neboť pro nám známé korpusy přirozeného jazyka se takovým jeví být.

Kritický čtenář se již jistě ptá, jestli můžeme data z tak omezeného intervalu (pro věty s průměrnou délkou tokenu od 3 do 5 písmen) zobecnit, avšak věty, které by měly průměrné tokeny kratší nebo delší, jsou vzácné, neboť tato vlastnost se distribuuje podle normálního rozdělení, a tak hodnoty oscilují kolem průměru. Tím je v arabském textu délka asi 4,4 písmene.

Důvod, proč na tento lineární vztah nikdo nepoukázal dřív, tkví zřejmě v tom, že k jeho změření je potřeba trochu neintuitivní abstrakce: nejprve setřídíme věty podle délky průměrného tokenu (v uvedeném grafu jsou věty setříděné od těch, jejichž průměrný token má 3 písmena po ty, jejichž průměrný token má 5 písmen, neboť těchto vět bylo nejvíc; jako jednotku jsem použil desetinu tokenu). Poté u každé skupiny vět určíme frekvenci průměrného slova. Získáme vztah, který můžeme jednoduše zapsat vzorcem (písmena jsem volil tak, aby se nepletla s výše uvedenými vzorci):

$$y = C\,x + b$$

kde y je průměrná četnost typů ve větě, jejíž tokeny mají průměrnou délku x. C a b jsou parametry, kde $c < 0$. A konečně formulovat závěrečnou hypotézu této kapitoly:

Hypotéza číslo 13: **Vztah průměrné četnosti typů ve větě k průměrné délce tokenů ve větě je lineární funkce se zápornou směrnicí.**

Tak máme před sebou tyto tři vzájemně související funkce:

frekvence slova = e (délka věty)

délka slova = f (délka věty)

frekvence slova = g (délka slova)

Nyní, abychom si ukázali v praxi, jak tyto tři vzorce souvisejí, můžeme si Altmannův vztah reformulovat:

délka slova = délka slova

délka slova = g' (frekvence slova)

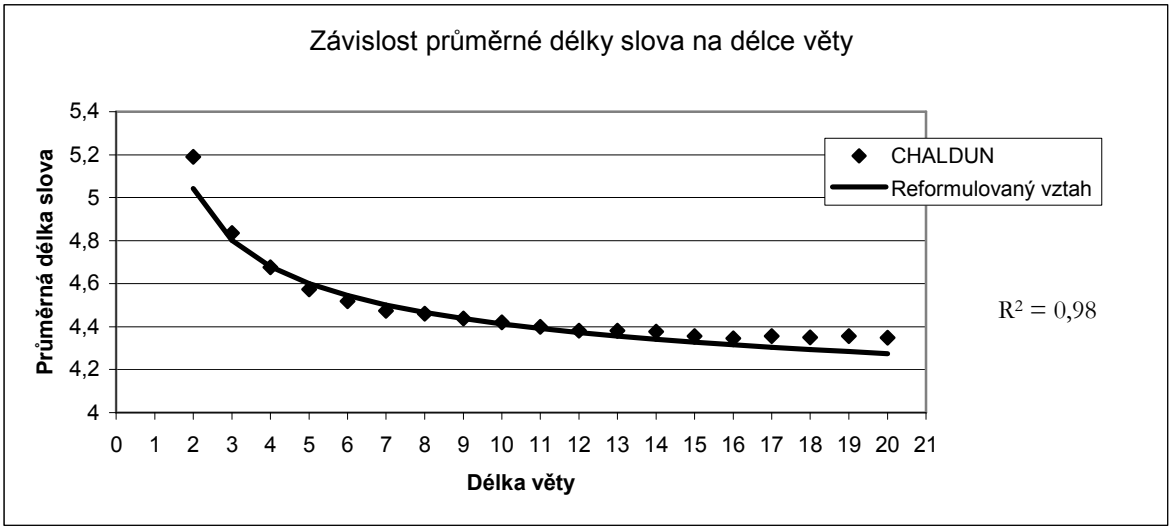
délka slova = g' (e (délka věty))

Tak získáme alternativní vzorec, který modeluje Menzerathův vztah:

$$y = \frac{k\ ln(ln(x)) + a - b}{C}$$

který používá parametry z příbuzných vztahů: *k* je přibližně rovno průměrné frekvenci typů ve větách o 15 slovech zmenšené o průměrnou frekvenci typů ve větách o 3 slovech; *a* je přibližně rovno frekvenci typů ve větách o 3 slovech, parametry *C* a *b* dosadíme ze vztahu průměrné četnosti typů na průměrné délce tokenů ve větě. (Jen připomínám, že *y* je průměrná délka slova o *x* tokenech).

Zkusme takto reformulovaný vztah, který používá parametry z jiných vztahů, uplatnit v praxi:



I když ani teď není model opticky uspokojivý, R^2 je dokonce vyšší než v původním vztahu[80].

Stejně můžeme reformulovat kteroukoli z funkcí *e*, *f* a *g*, takže při jejím výpočtu použijeme parametry ostatních dvou funkcí. Myslím, že podobným způsobem budeme schopni zřetězit mnoho dalších vztahů.

4.3 Shrnutí

V této kapitole uvádím tezi, která tvrdí, že „průměrná četnost slov ve větě stoupá s její délkou". Tento vztah se dá přesněji aproximovat rovnicí:

$$y = k \ln (\ln (x)) + a ,$$

kde *y* je průměrná frekvence typů ve větě o *x* slovech, *a* je přibližně rovno průměrné frekvenci typů ve větách o 3 tokenech a *k* + *a* je přibližně rovno průměrné frekvenci typů v textu (respektive průměrné frekvenci typů ve větách o 15 tokenech). Následně se ukazuje, že obdobnou rovnicí je možné aproximovat Menzerathův vztah aplikovaný na slova jako konstituenty a věty v roli konstruktů. Díky tomu nacházíme lineární vztah v závislosti průměrné četnosti typů na průměrné délce tokenů ve větě:

$$y = C x + b ,$$

kde *y* je průměrná četnost typů ve větě, jejíž tokeny mají průměrnou délku *x*. *C* a *b* jsou parametry, kde $c < 0$. Tak završujeme trojici funkcí, které vzájemně úzce souvisejí.

Popsané vztahy vznikly jako vedlejší produkt při řešení předchozích kapitol, přesto jsou možná nadějnější než jejich výsledky, ukazují totiž na propojení jednotlivých veličin, které můžeme na textu jednoduše naměřit, a na propojení různých úrovní jazyka. Předkládám pouze měření a modely, nikoli ucelenou teorii, která by se alespoň pokusila vysvětlit, proč tyto vztahy platí a jestli to má nějaké širší konsekvence. Rovněž preciznější testování, které je poměrně komputačně náročné, není v tuto chvíli jednoduché provést.

[80] To ovšem není žádné vítězství, neboť se zdvojnásobil počet parametrů, William Occam mi stojí za zády a brousí si břitvu.

5. Frekvence slova a jeho délka v nadvětných strukturách

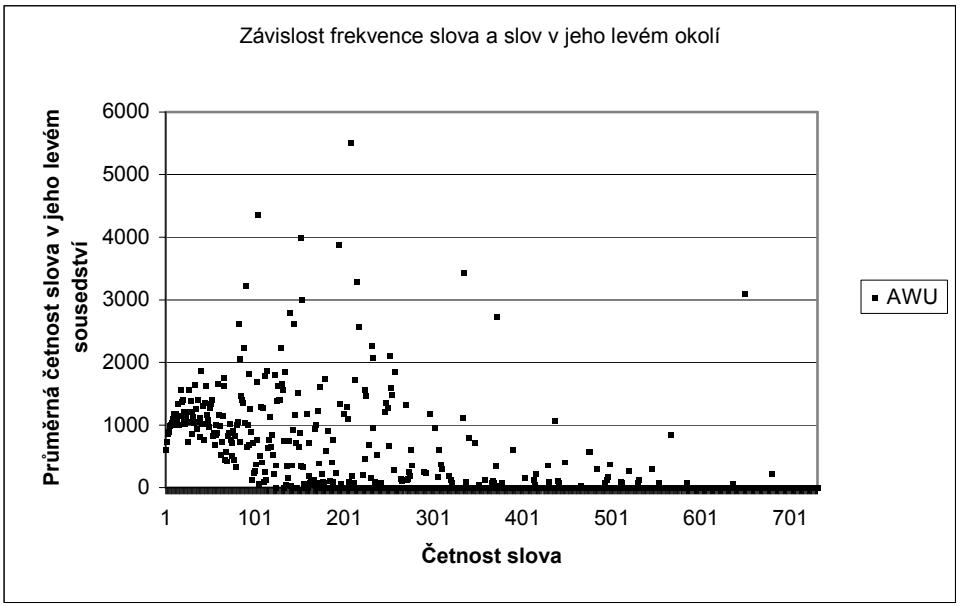
5.1 Frekvence slova a frekvence slov v jeho okolí

K jakému typu token náleží je do značné míry dáno tím, jaká slova se vyskytují v jeho okolí. Myslím, že na tom by se shodli lingvisté různých epoch a proveniencí, dokonce i zástupci protichůdných proudů. Tuto nezpochybňovanou skutečnost jsem se již pokusil kvantifikovat; výsledky se sice do této práce již nevejdou, nicméně doufám, že se ke čtenáři dostanou v některé příští publikaci.

Zeptejme se nyní trochu jinak: má frekvence slova vliv na to, jakou frekvenci bude mít slovo v jeho bezprostředním okolí? Náš výzkum frekvence slov ve větách napovídá, že by v četnosti slov mohly být nějaké přirozené rytmy. Zkusme je zachytit následujícím měřením:

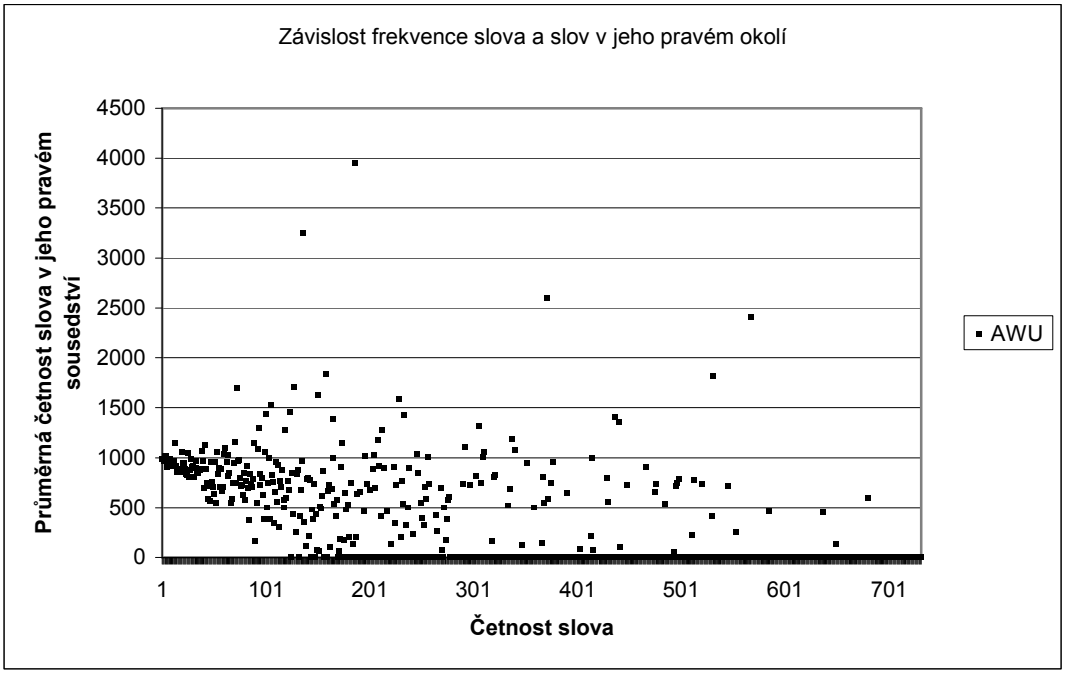
Slova daného textu roztřídíme do kategorií podle jejich frekvence (v daném textu). Poté pro každou kategorii zprůměrujeme frekvence slov, která bezprostředně sousedí s tokeny v oné kategorii[81].

Vyneseme do grafu závislost průměrné frekvence typů, které bezprostředně sousedí s daným slovem, na frekvenci tohoto slova[82]:

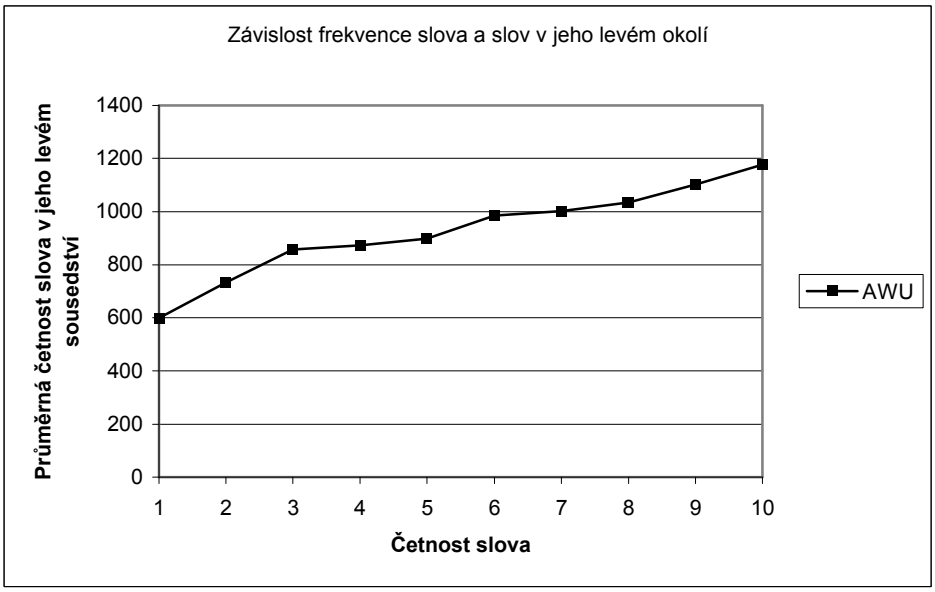


[81] Příklad měření naleznete v Příloze A.

[82] Měřeno na korpusu AWU. Ačkoli se jedná o arabský text, pro jednoduchost pod pojmem levé okolí myslíme slovo, které danému slovu předchází, pod pojmem pravé okolí myslíme slovo, které po daném slovu následuje.



Vidíme, že data jsou značně chaotická, což mě demotivuje od podstupování dalších měření v této oblasti a vede ke smíření se s tím, že pokud podobná závislost existuje, bude složitější se k ní propracovat. Určitý náznak vidíme na začátku grafu pro levé okolí – vidíme, že málo četná slova (a těch je také nejvíc, vzpomeňme na zipfovskou distribuci, takže zde jsou data nejspolehlivější) často předcházejí také méně četná slova:

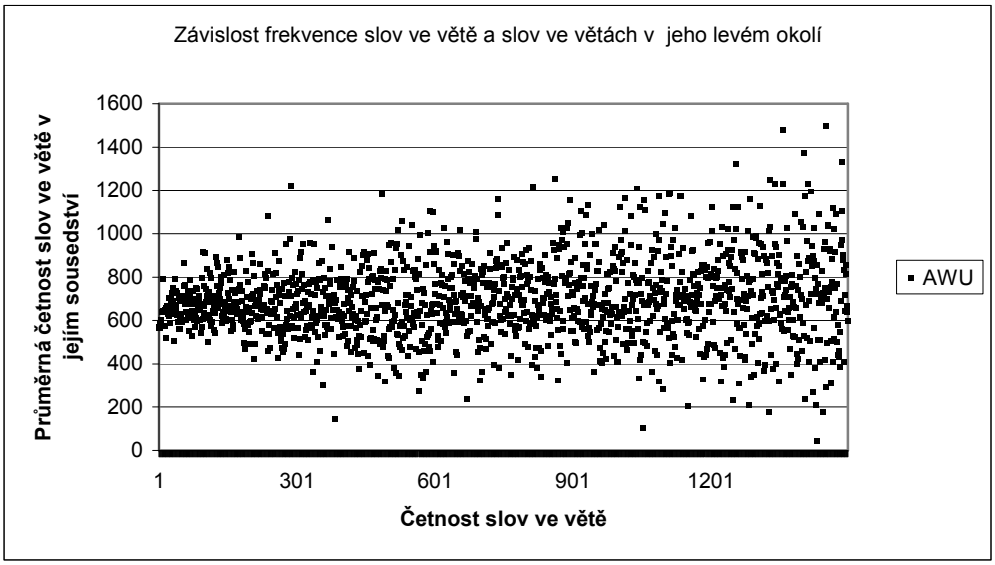


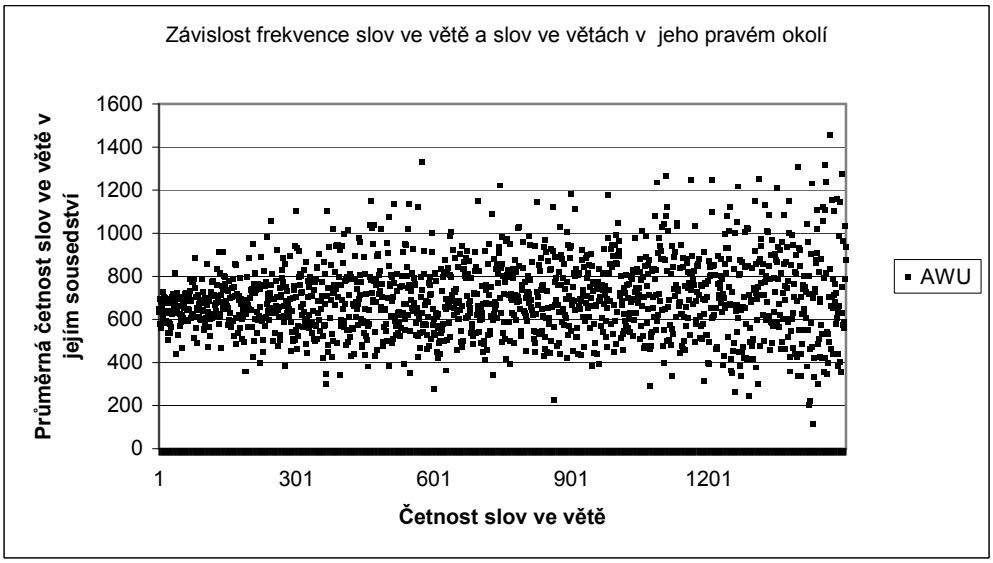
Tento graf nás nabádá k vyslovení hypotézy číslo 14:

Frekvence slova je přímo úměrná průměrné frekvenci slova, které mu předchází.

Ovšem slov s vyšší frekvencí než deset je celkem v textu málo, a tak se zvyšuje rozptyl dat. Proto hypotézu zatím označme jako netestovanou a nechme si ji do budoucna. Její platnost by znamenala, že slova s určitou frekvencí mají tendenci vytvářet jakési klastry.

V této souvislosti prověřme, jak se v obdobném testu budou chovat celé věty. Nebudeme testovat frekvenci vět, ale průměrnou četnost slov, které obsahují:





Grafy neukazují na to, že by mezi průměrnou frekvencí slov ve větě a průměrnou frekvencí slov ve větách, které s ní sousedí, byla nějaká souvislost.

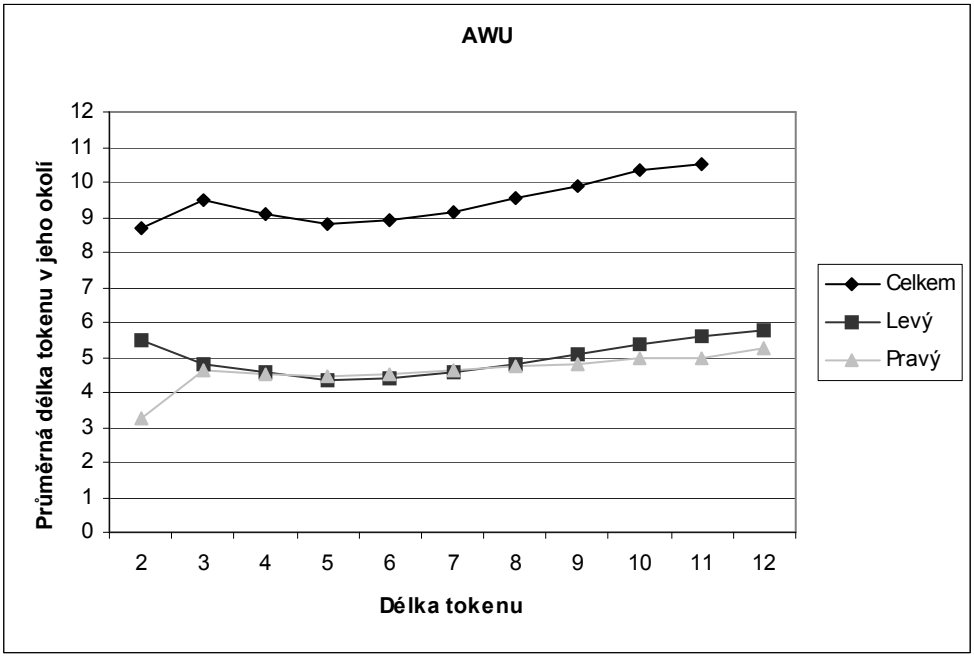
Tyto výsledky uvádím nejen proto, že ukazují směr (slepou uličku) mým následovníkům, ale také pro úplnost, neboť podobná měření provedeme v následující kapitole i pro délky slov a vět. Vztahy, které jsme formulovali v předchozí kapitole, nás nepochybně svádějí k tomu, abychom obdobné závislosti jako na poli délky, hledali i na poli frekvence, ovšem, jak si ukážeme v zápětí, analogie není dokonalá.

5.2 Délka slova a délka slov v jeho okolí

Analogicky k předchozí kapitole provedeme následující měření:

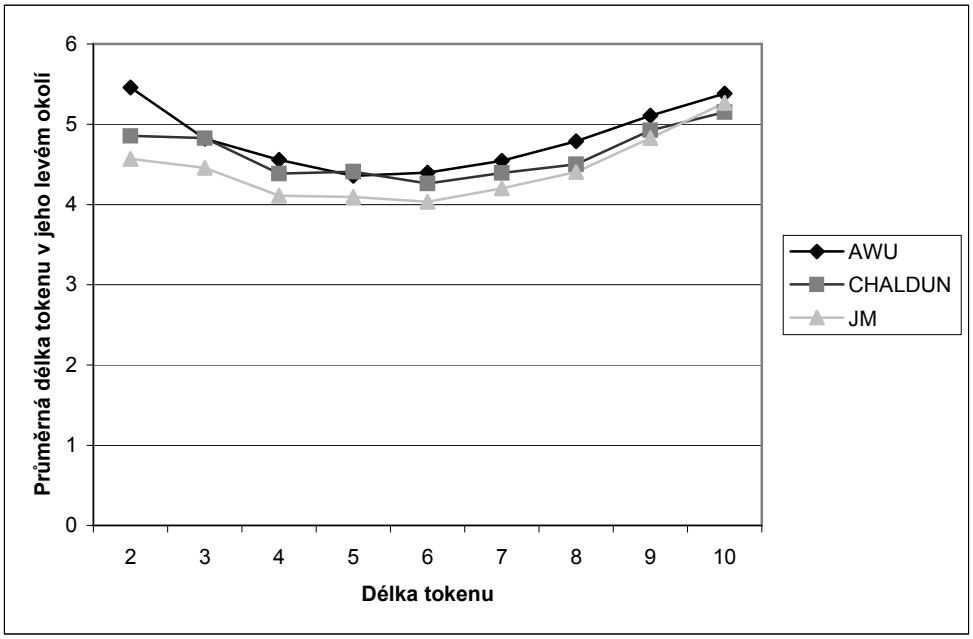
Tokeny daného textu roztřídíme do kategorií podle délky. Poté pro každou kategorii zprůměrujeme délky tokenů, které bezprostředně sousedí s tokeny v oné kategorii[83].

Vztah mezi frekvencí slova a frekvencí jeho sousedů nebyl přesvědčivý. Proto nás možná překvapí výsledky této kapitoly, ve které si ukážeme, že vztah mezi délkou tokenu a délkou tokenů v jeho bezprostředním okolí je poměrně zřetelný. Pokusme se ho napřed vizualizovat pomocí grafu:

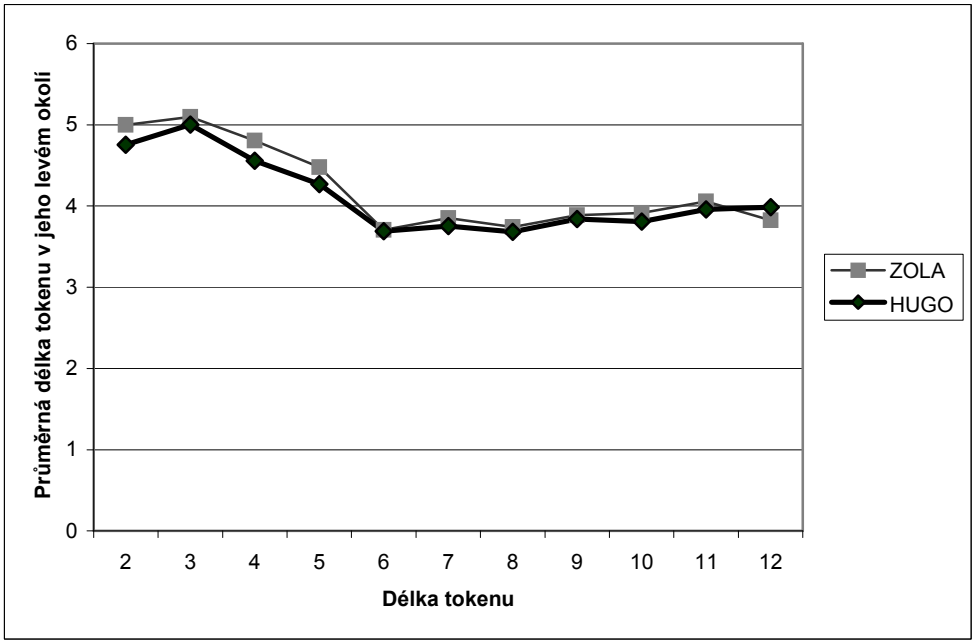


Délka tokenů se sice také nedistribuuje podle normálního rozdělení, ale výkyvy nejsou tak velké jako v případě frekvence, díky tomu se dostáváme k poměrně hladké křivce. Jak ji charakterizovat? Netroufám si matematicky ji modelovat, avšak slovně ji můžeme popsat tak, že graf pro pravý kontext vypadá jako mírně rostoucí, křivka pro levý kontext je zajímavější – ze začátku se délka průměrného kontextu snižuje s rostoucí délkou tokenu, přibližně po tokenu o pěti písmenech průměrná délka levého souseda začíná naopak růst. Jak tato závislost vypadá v ostatních korpusech?

[83] Příklad měření naleznete v příloze A.



Arabské korpusy CHALDUN, JM a AWU opět přes tisíciletou propast vykazují shodné rysy. Naměřené křivky se však liší od těch, které jsme nalezli u jiných jazyků:



Zjevně jsme opět narazili na vztah, ve kterém křivka není jednoduše popsatelná, nicméně pro texty napsané v jednom jazyce je podobná. Podívejme se, jak jsou si podobné jednotlivé závislosti délky tokenu na délce tokenu v jeho **levém okolí** na jednotlivých korpusech (Pearsonův koeficient, tokeny o délce 2 – 10 písmen):

	AWU	COOPER	MLOCI	NAHODNY	JM	CHALDUN	HUGO	ZOLA
AWU	**1,0000**	*0,0057*	0,6650	0,6238	0,8995	**0,9027**	*0,0682*	*0,0646*
COOPER		**1,0000**	*0,0155*	*0,1260*	*-0,0311*	*0,1131*	0,8075	0,7488
MLOCI			**1,0000**	*0,4972*	*0,2993*	*0,3691*	*0,4066*	*0,4290*
NAHODNY				**1,0000**	0,5068	0,6652	*0,2271*	*0,2480*
JM					**1,0000**	**0,9250**	*-0,1613*	*-0,1757*
CHALDUN						**1,0000**	*0,0280*	*0,0097*
HUGO							**1,0000**	**0,9933**
ZOLA								**1,0000**

Podobně jako v kapitole 2.4 a 3.4 najdeme vysoký korelační koeficient (kolem 99 %) pro vztah mezi korpusy ZOLA a HUGO, také datové řady pro tři arabské korpusy CHALDUN, JM a AWU mají mezi sebou vysokou korelaci – kolem 90 %. Mezi daty z ostatních korpusů je korelace obdobná, jako pro náhodný text.

Tady máme tabulku pro **pravý kontext** (Pearsonův koeficient, tokeny o délce 2 – 10 písmen):

	AWU	COOPER	MLOCI	NAHODNY	JM	CHALDUN	HUGO	ZOLA
AWU	**1,0000**	*-0,6982*	*-0,8013*	*-0,4819*	0,6384	*0,4856*	*-0,5706*	*-0,5380*
COOPER		**1,0000**	0,5810	*0,3430*	*-0,4964*	*-0,2360*	0,8376	0,8822
MLOCI			**1,0000**	0,7045	*-0,0989*	*0,0347*	*0,4021*	*0,3755*
NAHODNY				**1,0000**	*-0,0975*	*0,4451*	*0,2478*	*0,2058*
JM					**1,0000**	0,6955	*-0,4923*	*-0,4675*
CHALDUN						**1,0000**	*-0,2241*	*-0,2193*
HUGO							**1,0000**	**0,9890**
ZOLA								**1,0000**

Vidíme, že pravý kontext se chová mnohem chaotičtěji než levý – kromě vztahu mezi korpusy ZOLA a HUGO se datové řady naměřené na korpusech přirozeného jazyka chovají podobně jako řady naměřené na náhodně zpřeházeném textu. Opatrně formulujeme další hypotézu:

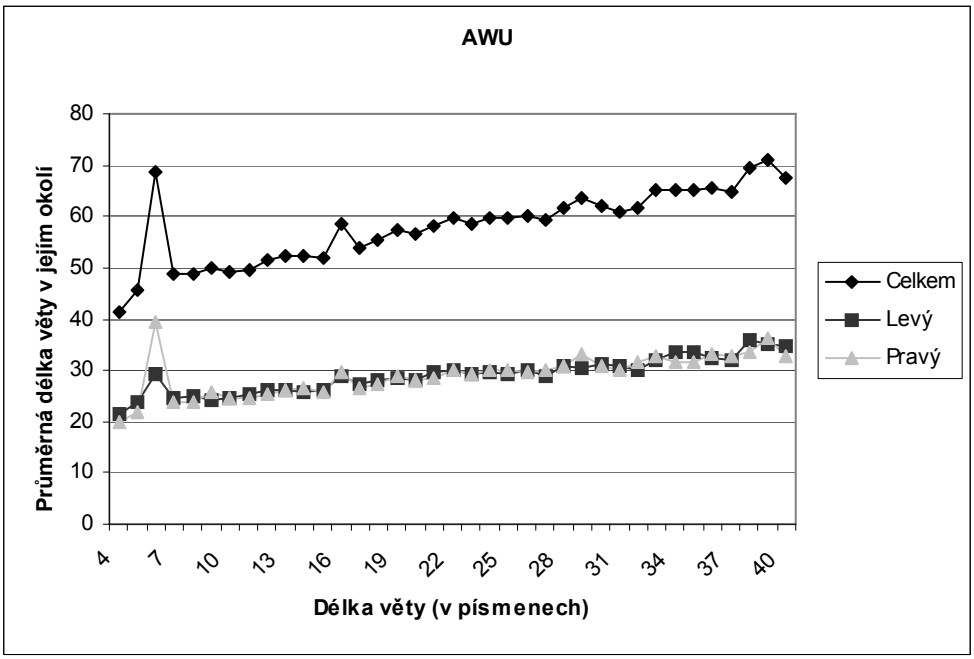
Hypotéza číslo 15: **Jsou-li texty napsány ve stejném jazyce, pak existuje významná korelace mezi vztahy délky tokenů a průměrnou délkou tokenů v jeho levém okolí pro tyto dva texty.**

Výše uvedenou tabulku chápejme jako testování této hypotézy.

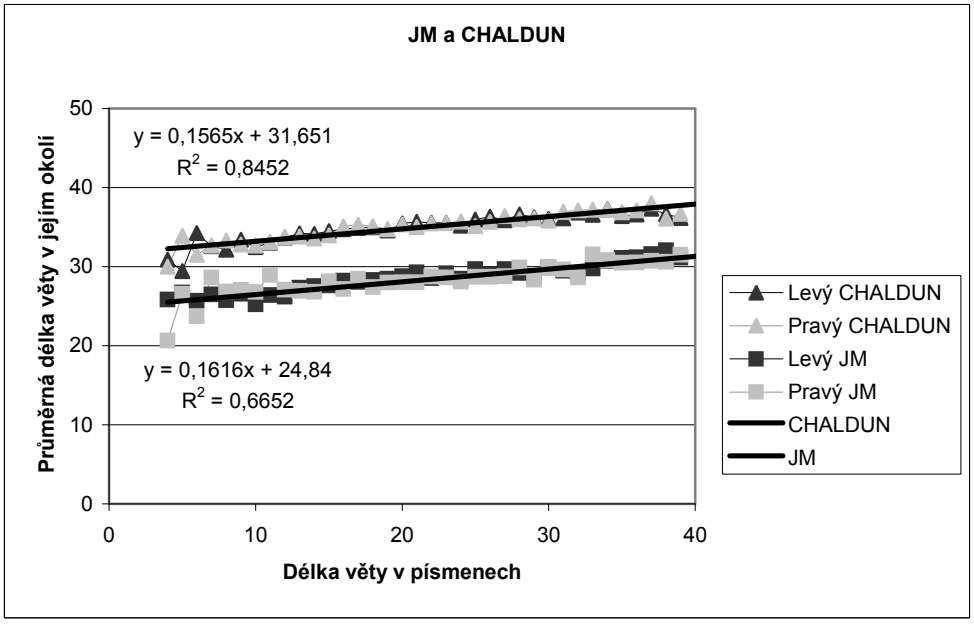
Tato závislost platí i přes hranice vět, pokud ovšem chceme zkoumat nadvětné struktury, měli bychom se podívat, jak se budou v obdobném vztahu chovat celé věty.

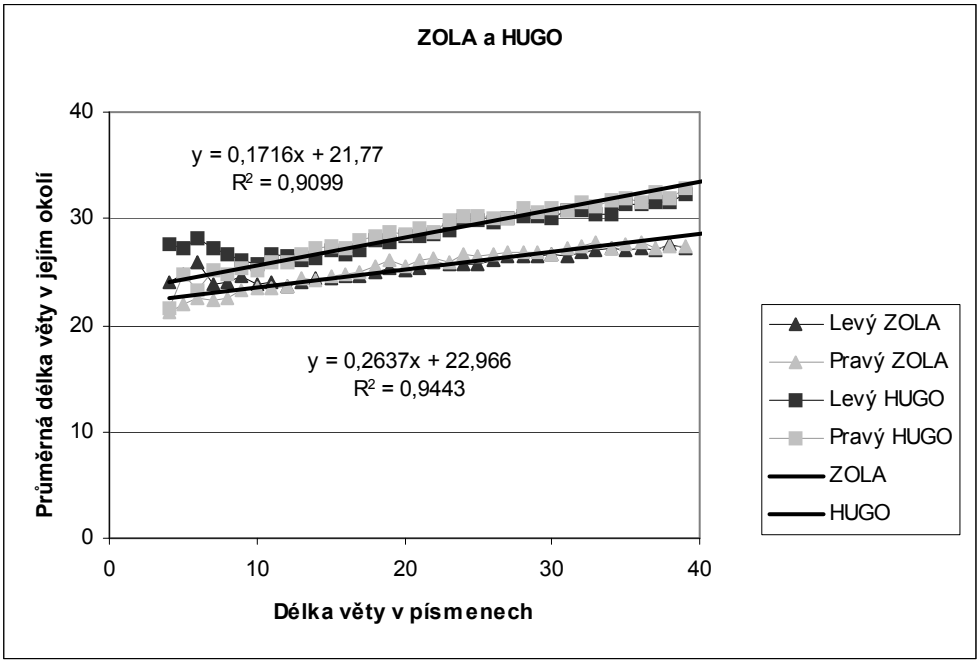
Nejprve roztřídíme věty daného textu do kategorií podle délky. Poté pro každou kategorii zprůměrujeme délky vět, které bezprostředně sousedí s větami v oné kategorii.

Tady je ukázkový graf pro tuto závislost:



Na první pohled si všimneme lineárního vztahu. Zároveň hlavním rozdílem od předchozí závislosti je, že datové body pro bezprostřední levé okolí jsou téměř totožné s těmi pro pravé bezprostřední okolí. Tento graf se dá nejspíše interpretovat tak, že se v textu vytvářejí jakési shluky krátkých a naopak dlouhých vět. Je tento vztah lineární i pro jiné korpusy?

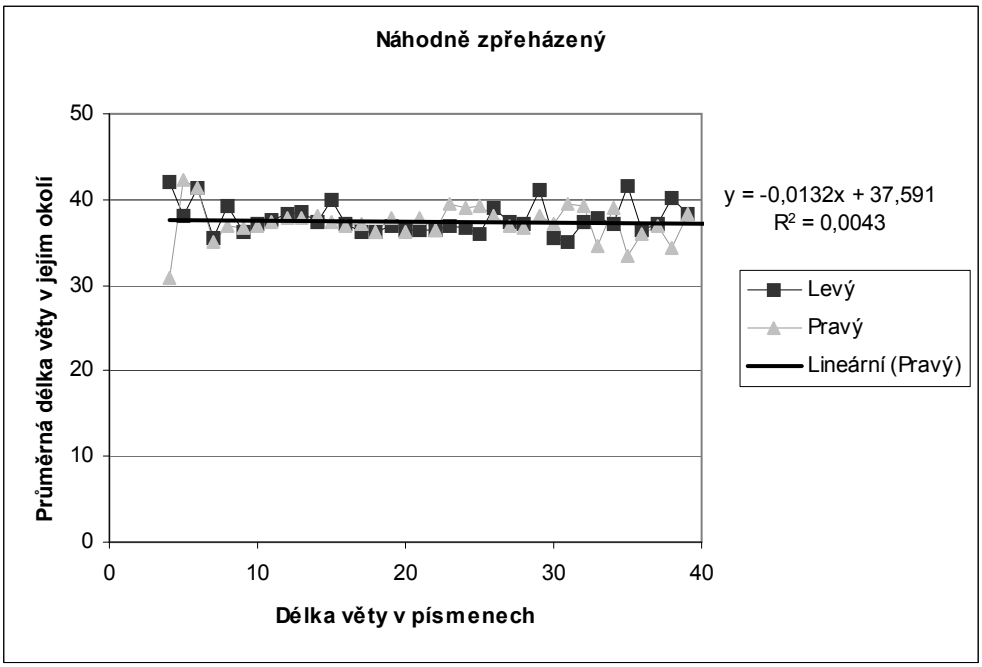




Pro dva arabské středověké korpusy JM a CHALDUN je daný vztah charakterizovatelný jako ukázněná lineární závislost, směrnice regresí pro oba korpusy je obdobná a data pro levý i pravý kontext se od sebe výrazně neliší (podobně je tomu u korpusu MLOCI i COOPER). Naproti tomu ZOLA a HUGO mají směrnice na první pohled vzájemně odlišné a dokonce se liší i data naměřená na levém a pravém kontextu. Křivka, kterou jsme naměřili pro levý kontext korpusu HUGO, se podobá té pro délku slov ze začátku této kapitoly – také napřed klesá a následně stoupá (přelomová je věta o deseti písmenech). Tím, že obdobná data získáváme pro oba korpusy, nemůžeme tuto anomálii svést na náhodné odchylky, možná zde hraje roli hrubá chyba při měření, věty o několika málo písmenech jsou poněkud podezřelé, i když je jich v textu hodně. Tyto dva francouzské korpusy jsou z naší kolekce jediné, které jsem nekontroloval, a jazyková nekompetentnost mi brání, abych zjistil, v čem jsou krátké věty v těchto korpusech odlišné. Protože se naše práce týká arabského textu, na kterém vztah fungoval velmi dobře, můžeme si dovolit zúžit platnost naší hypotézy pro arabštinu:

Hypotéza číslo 16: **Pokud je text napsán v arabském jazyce, pak průměrná délka vět v pravém i levém bezprostředním okolí vět, které jsou dlouhé *n* písmen, je přímo úměrná *n*.**

Výše uvedené grafy mohou sloužit jako pokus o vyvrácení této hypotézy. Pro úplnost si ještě uveďme, jak se vztah chová na korpusu s náhodně transponovanými tokeny (včetně náhodně transponované interpunkce):



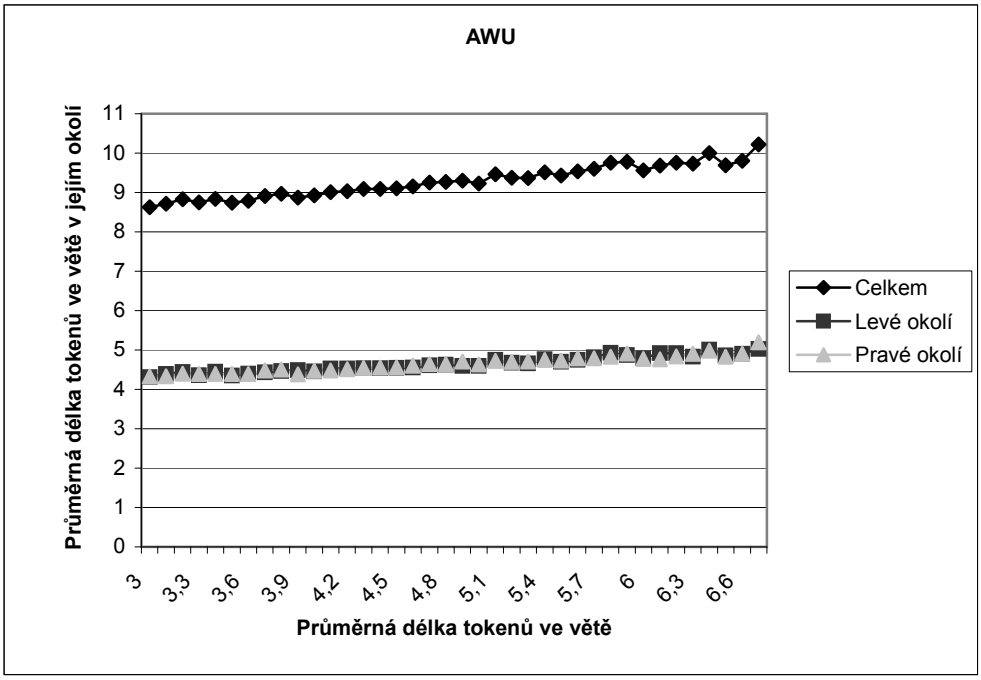
Vidíme, že data naměřená na „textu“ s náhodně zpřeházenými tokeny se podle tohoto zákona nechovají, čili že se (podle očekávání) jedná o specifickou vlastnost *textu*.

Doufám, že se v budoucnu dostanu k tomu, abych zjistil, proč vztah neplatil pro všechny korpusy, a zobecnil jej pro všechny jazyky.

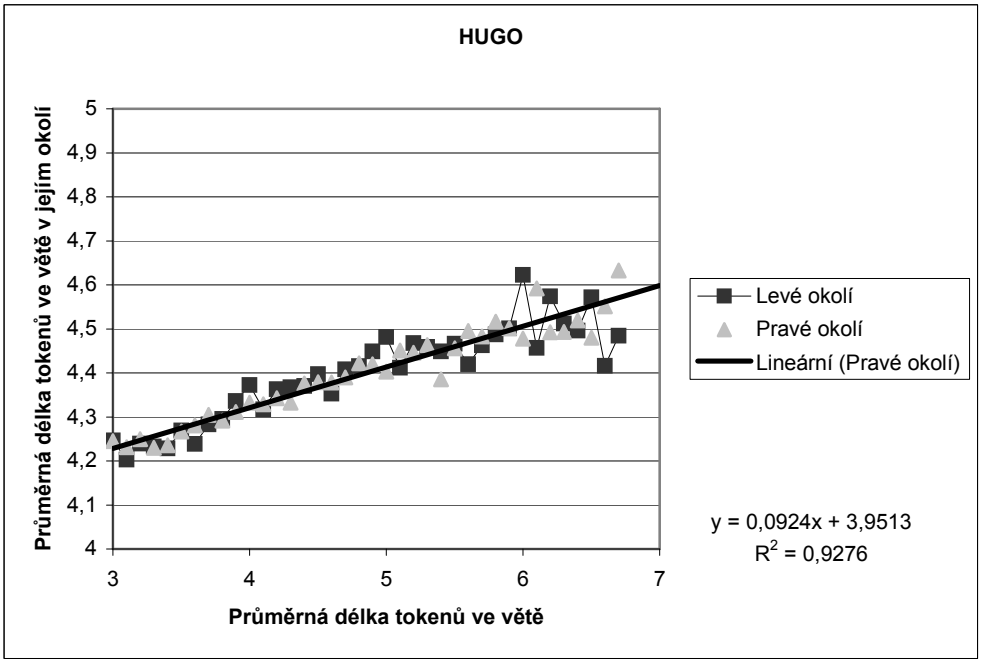
Tuto lineární závislost můžeme interpretovat tak, že mluvčí střídají úseky textu s relativně krátkými větami a úseky s větami relativně dlouhými, což se vcelku shoduje s tím, co intuitivně cítíme při četbě. Nyní se zaměříme na ještě jeden obdobný vztah, který se nám nabízí, podobně jako v kapitole 5.1:

Nejprve roztřídíme věty daného textu do kategorií podle průměrné délky slov, které obsahují. Poté pro každou kategorii zprůměrujeme délky slov, která obsahují věty, které bezprostředně sousedí s větami v oné kategorii.

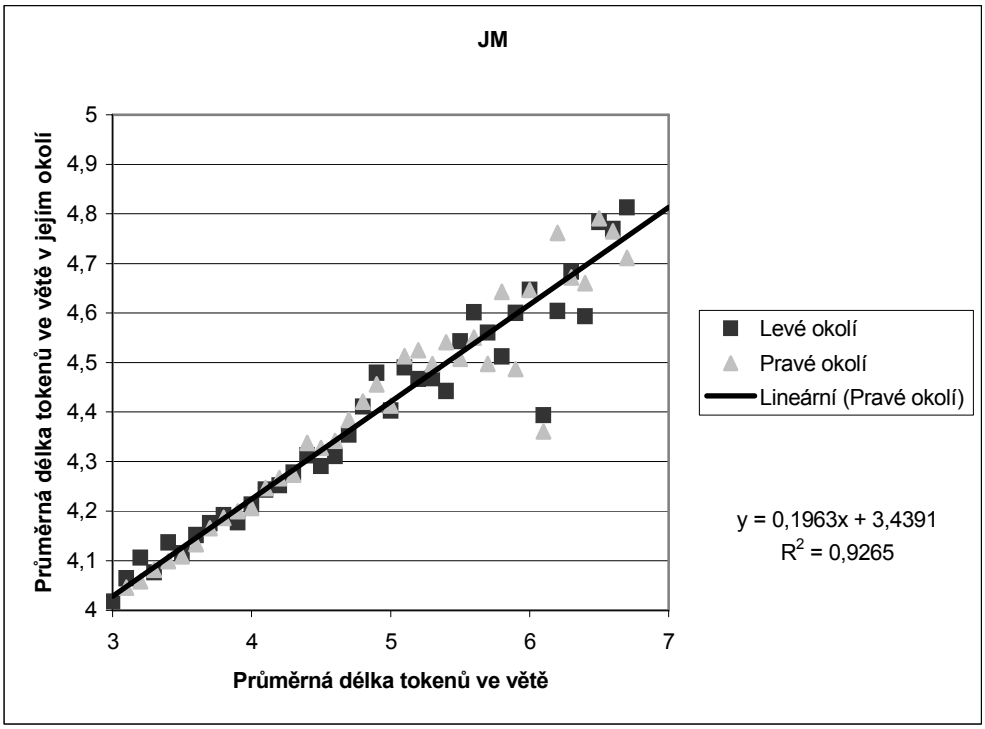
Získáme datovou řadu jako je například tato, kterou jsme naměřili na korpusu AWU:



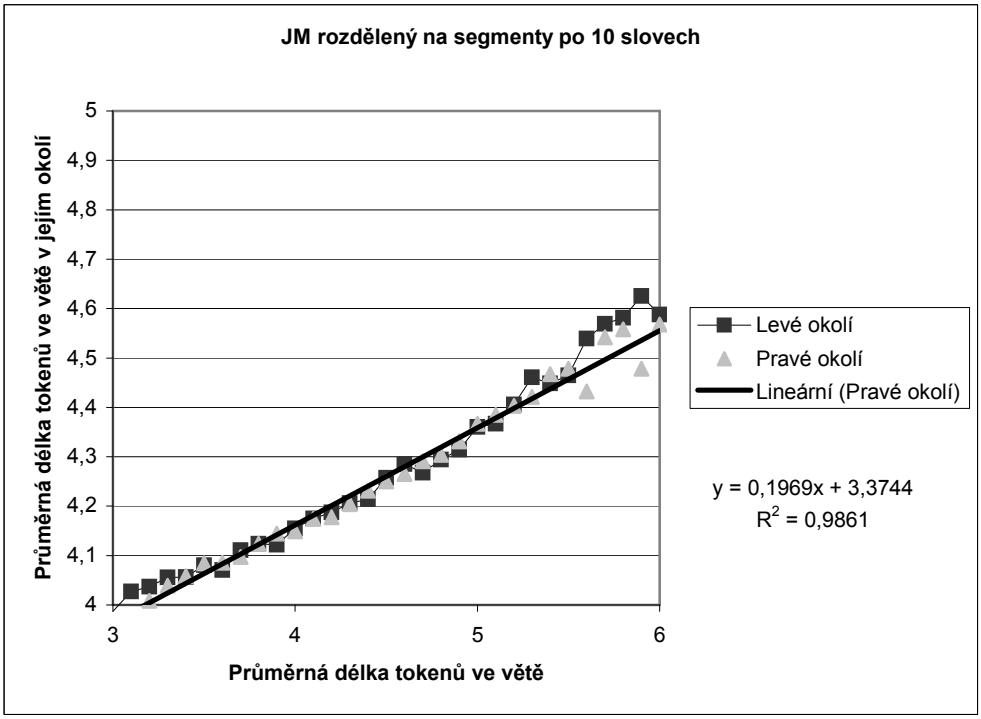
Graf ukazuje na krásnou lineární závislost[84] a naše interpretace tohoto jevu bude také přímočará: Věty, které obsahují relativně krátká slova, často sousedí s větami, které obsahují také relativně krátká slova. A naopak. Podobně vypadá tato závislost i v jiných korpusech, ať už arabských, či nearabských:

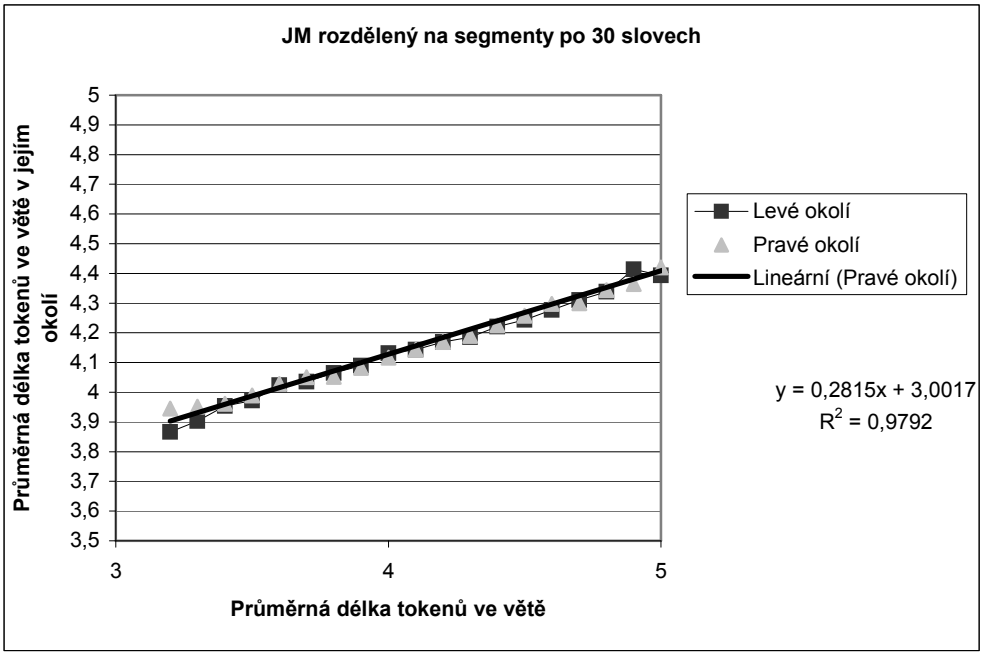


[84] Neodpustil jsem si toto hodnotící adjektivum, i když neexistuje žádný důvod, proč by lineární vztahy měly být lepší, než jiné křivky. Nicméně běžný člověk, když se dopracuje k přímé úměře, má jaksi pocit, že danému jevu rozumí, v ostatních případech se pídí po dalším vysvětlení. Spousta lidí strávila množství času, aby osvětlila, proč se slova řídí Zipfovským rozdělením (včetně samotného Zipfa), kdyby byl Zipf nalezl lineární vztah, nejspíš by k filozofování nad tímto tématem příliš neinspiroval.



Otázkou ovšem je, jestli je vhodné formulovat tento vztah pouze pro věty a není výhodnější jej rovnou zobecnit pro jakékoli, třeba i mechanicky rozdělené celky textu. Když totiž korpus JM rozdělíme na segmenty po deseti a třiceti slovech, zjistíme, že nalezneme obdobný vztah:

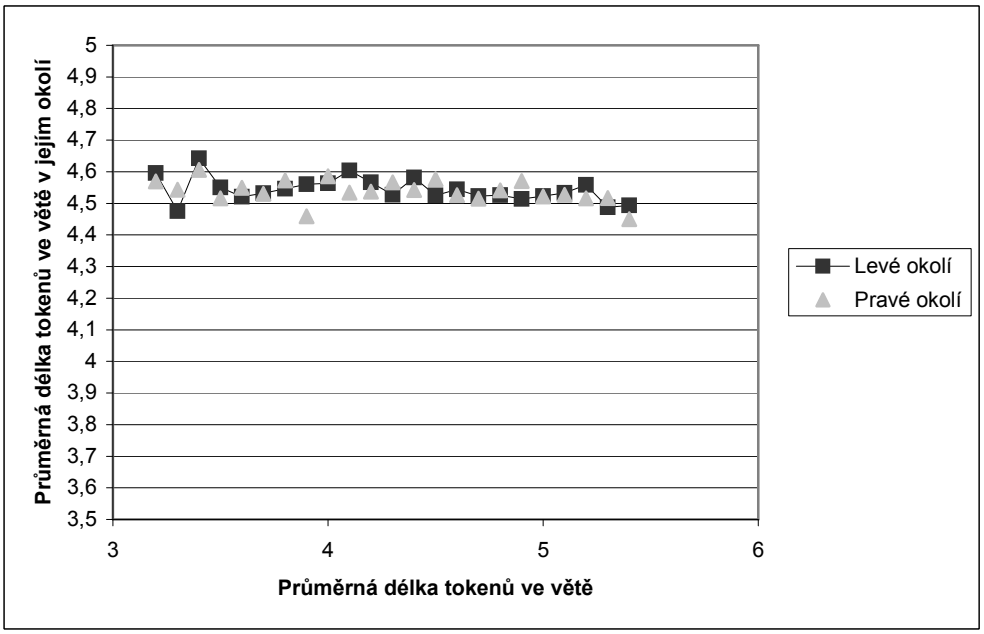




Vypadá to tedy, že zde nahlížíme do struktur, které s větami nemají nic společného a že následující hypotéza se bude týkat samotných tokenů; jde tady spíše o jakési rozšíření hypotézy ze začátku této podkapitoly:

Hypotéza č. 17: **Průměrná délka slov ve větách (segmentech) v pravém i levém bezprostředním okolí vět (segmentů), které obsahují slova o průměrné délce *n* písmen, je přímo úměrná *n*.**

Ani tato hypotéza – podle očekávání – neplatí pro „text“ s náhodně zpřeházenými tokeny:



Tato hypotéza vzbuzuje množství otázek: Závisí nějak směrnice tohoto lineárního vztahu na délce segmentů? Jak dlouhé (krátké) segmenty ještě mohou být, aby tento zákon vůbec ještě platil? Při zachování stejné délky tokenů, jsou parametry rovnice typické pro jazyk, nebo pro styl? Šetření, která jsem provedl, napovídají, že jejich zodpovězení nebude jednoduché.

Pokud by někdo viděl nějaký užitek v dalším výzkumu na tomto poli, bylo by jednoduché měřit například vztahy mezi délkou věty a průměrnou délkou slov ve větách v jejím okolí, nebo třeba závislost mezi průměrnou frekvencí slov ve větě a průměrnou délkou vět v jejím okolí… Takových vztahů najdeme nespočet a pomocí hypotéz kapitoly 4 bychom mohli odvozovat čistě algebraickou cestou parametry jedné závislosti z parametrů závislosti jiné. Je otázkou, jestli bychom se tak přiblížili k ucelenějšímu pochopení zákonitostí, které v jazyce panují.

5.3 Frekvence a délka slova jako sémantický ukazatel

Mnohý čtenář možná nabyl dojmu, že téma frekvence slova jsme již plně vyčerpali, nicméně ještě se u ní zastavíme, neboť nám pomůže jako můstek, který bude spojovat předchozí a následující kapitolu.

Prozatím jsme se dívali na ty vlastnosti a zákonitosti frekvenčního zobrazení, které byly společné pro všechny texty v témže jazyce, popřípadě ukazovaly na shodné rysy všech jazyků. Nyní se naopak zastavme u myšlenky, jestli není možné pomocí frekvence a délky zachytit obsah textu, jestli nenajdeme veličinu, která by zůstávala stejná nebo podobná pro různé jazyky na textech s podobným významem (přeložených textech). Tato myšlenka najde oporu při našem intuitivním vnímání textu – například dramatická pasáž vyžaduje jasná, hodně frekventovaná slova a to nezávisle na jazyce, naopak při meditativní pasáži autor bude používat méně častá slova a tento rys chtě nechtě bude muset překladatel přenést i do jinojazyčné verze.

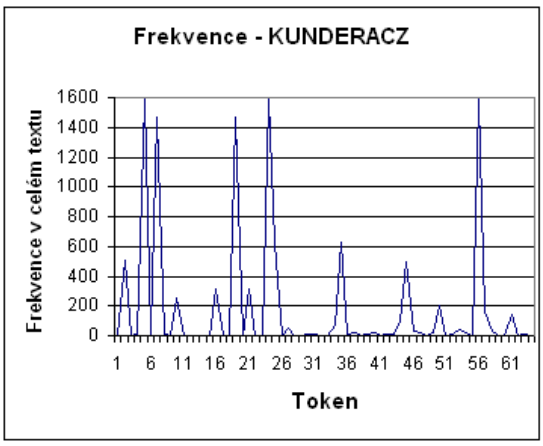
Obyčejné frekvenční a délkové zobrazení, jak jsme si ho definovali v předchozí kapitole, je příliš chaotické, než abychom mohli porovnávat přímo. To si můžeme názorně ukázat na úryvku Kunderova Valčíku na rozloučenou a jeho arabském překladu[85]:

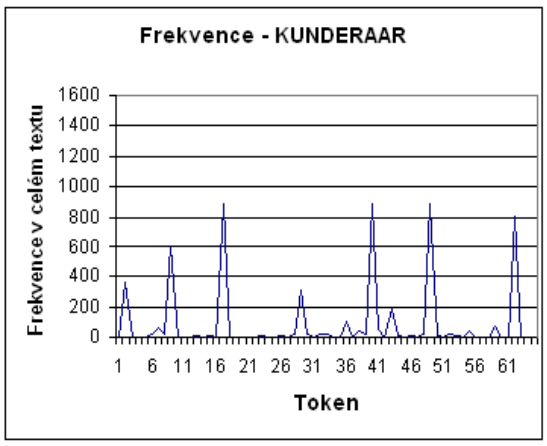
je středa ráno a lázně se opět probudily k čilému životu proudy vod crčí do van maséři se opírají do obnažených zad a na parkoviště právě přijelo osobní auto nikoli luxusní limuzína která na stejné místo přistála včera nýbrž obyčejné auto jaké má v této zemi většina lidí za volantem seděl muž asi pětačtyřicetiletý a byl sám zadní sedadla byla zatarasena několika kufry

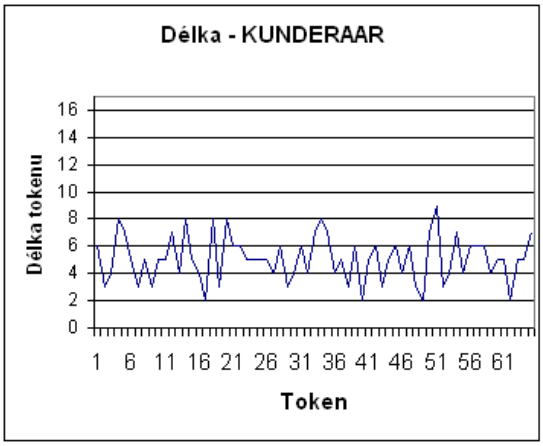
كان صباح الاربعاء واستيقظ النبع مرة ثانية على دائرة نشاطه المنهمك بدات انبثاقات الماء تسري فى الانابيب ثنى المدلكون اذرعهم ومفارش جديدة تتجهز ووقفت سيارة خاصة مباشرة الى مكان الجراج ليست السيارة المكشوفة الفارهة التى احتلت نفس المكان فى اليوم السابق لكن سيارة مظه

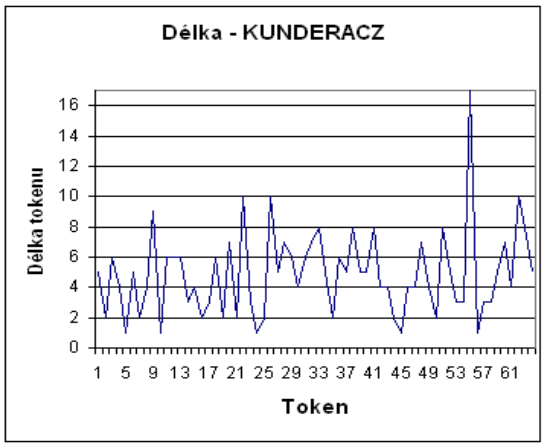
Grafy naměřené na tomto úryvku vypadají takto:

[85] Zlomky korpusů KUNDERAAR a KUNDERACZ, více o nich viz kapitolu 9. Arabský zlomek je překladem českého zlomku.









Grafy pro český a arabský text si nejsou příliš podobné. Není ani příliš divu, neboť české tokeny, tak jak následují jeden po druhém, neodpovídají arabským tokenům. Můžeme však využít vlastnosti, kterou jsme nalezli v podkapitolách 5.1 a zejména 5.2, totiž že tokeny s nízkou frekvencí, respektive malou délkou – a naopak, tvoří jakási delší pole. To by znamenalo, že v delším horizontu bychom měli vidět mírně stoupající a klesající křivky, a tak bychom si mohli položit otázku, jestli si tyto křivky pro přeložené texty odpovídají. Klouzavý průměr mnoho neřeší, pokusíme se tedy vymyslet lepší způsob: od datové řady odečteme její průměr. Následně každému datovému bodu *p* přiřadíme hodnotu součtu datových bodů od začátku po datový bod *p*. Poněkud jasněji tento postup vyjádříme vzorcem:

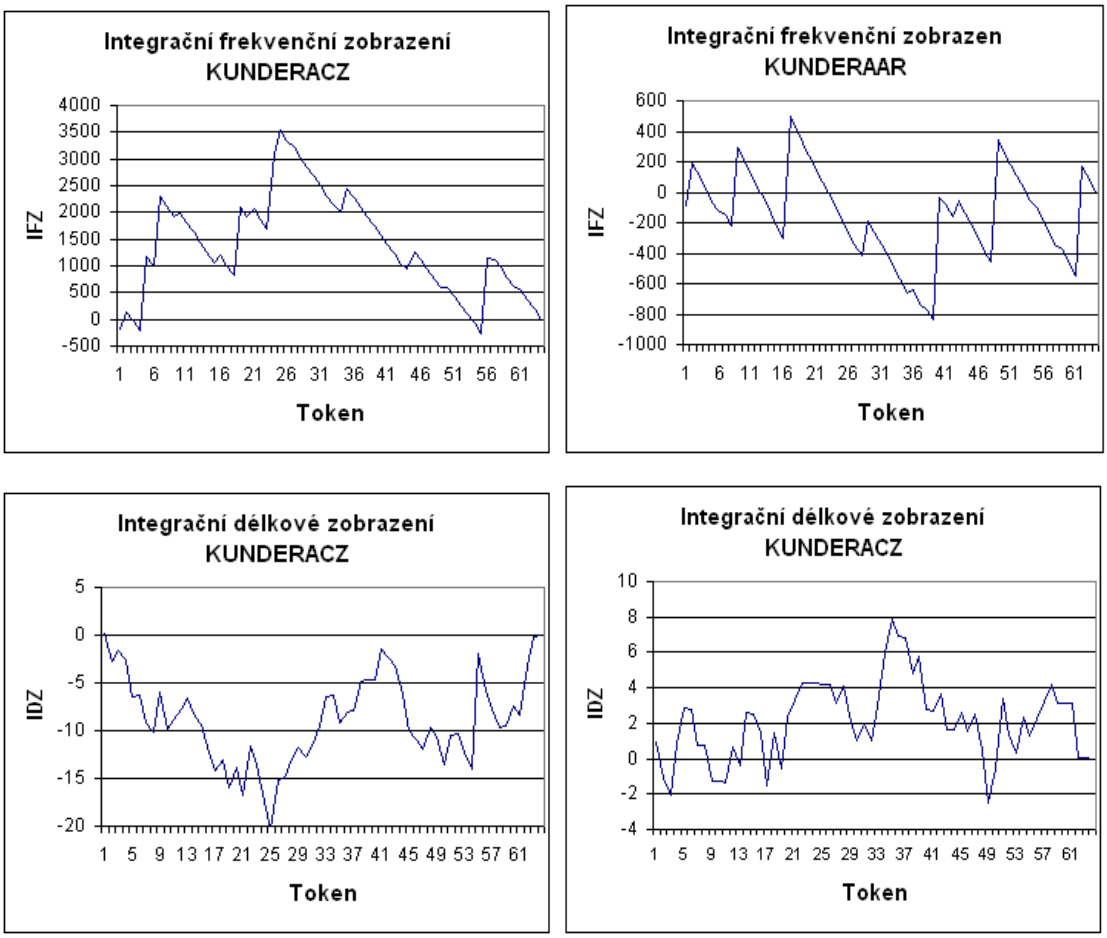
$$y_p = \sum_{i=1}^{p} (x_i - \bar{x})$$

V rekurentním tvaru vypadá následovně:

$$y_1 = x_1 - \bar{x}$$

$$y_p = y_{p-1} + x_p - \bar{x}$$

Tento postup si nazvěme jako **integrační zobrazení**[86]. Jeho základní vlastností je, že začíná i končí na hodnotě x_l – průměrné x a pokud bychom ho chtěli interpretovat, pak rostoucí křivka indikuje, že v původní datové řadě byly na tomto místě nadprůměrně velké hodnoty a naopak. A rovnou se pojďme podívat, jaké možnosti nám nabízí[87]:



Skutečně vidíme, že se grafy vzájemně trochu podobají (ani v jednom případě není podobnost zrovna očividná, nicméně při troše fantazie nalezitelná; při použití delšího textu se bude zvětšovat). Jak ale porovnávat různě dlouhé datové řady? V kapitole 2.4 jsme zmínili, že je možné změnit délku datové řady podle nejbližšího členu. Tento postup není jediný a asi ani ideální, nyní ho však pro jeho jednoduchost použijeme[88]. Rozpočítáme datové řady na 10000 položek[89]:

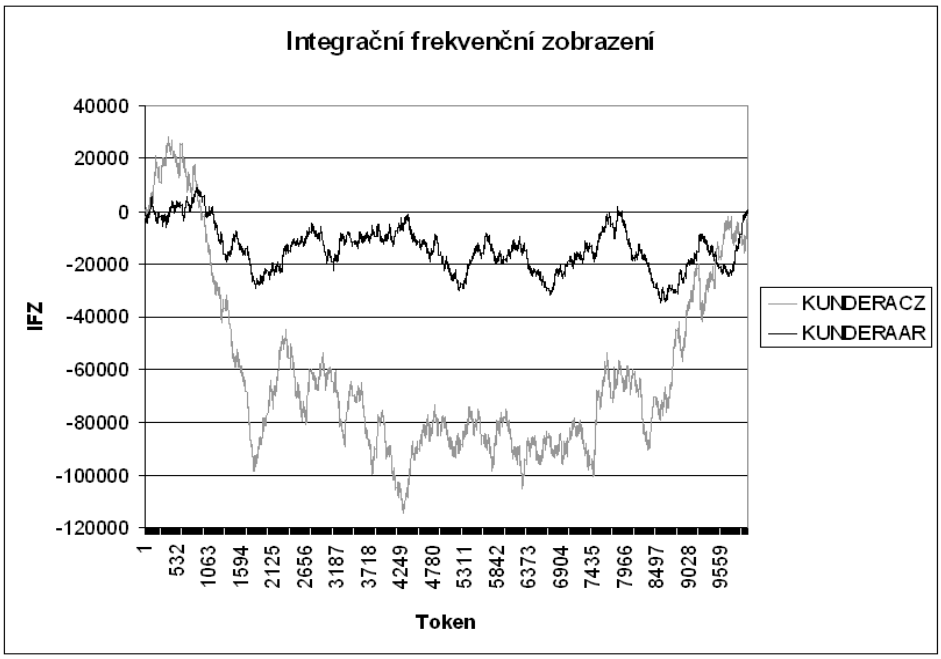
---

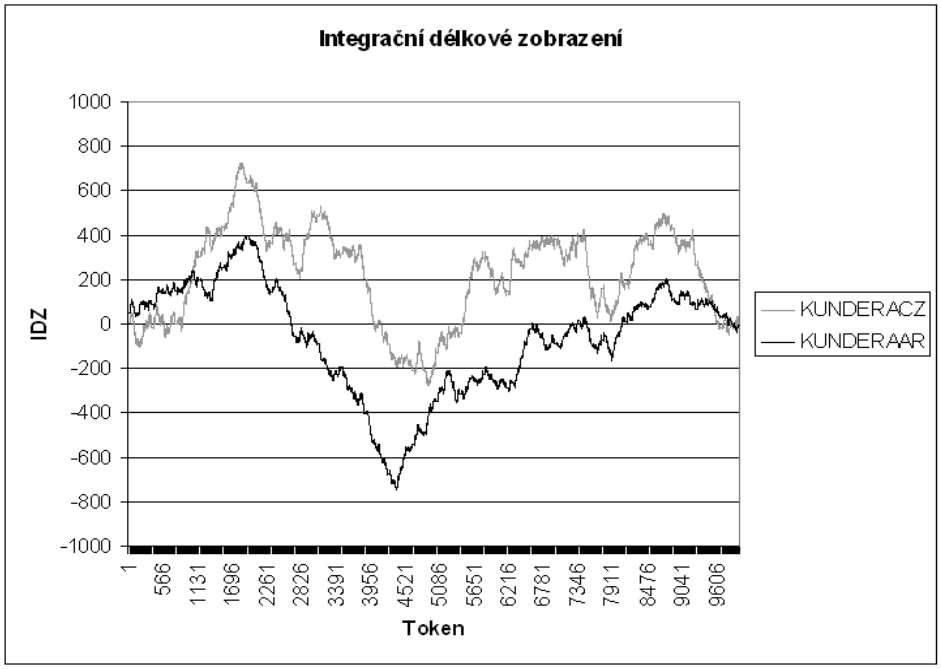
[86] I když zde s integrálem vlastně nepracujeme, a tedy nebude špatné, pokud se pro tuto funkci časem najde lepší pojmenování. Pro integrační zobrazení délkového zobrazení (zjednodušeně integrační délkové zobrazení) zavádím zkratku IDZ, pro integrační zobrazení frakvenčního zobrazení IFZ.

[87] Změřeno na stejných zlomcích korpusů KUNDERAAR a KUNDERACZ

[88] Aplikujeme poněkud modifikovaný algoritmus. Pojďme se na něj nyní podívat blíže:

Máme datovou řadu g_1 až g_m transformovat na datovou řadu h_1 až h_n. Použijeme tento vzorec:





Nyní můžeme použít Pearsonův korelační koeficient; pro integrační délkové zobrazení mezi českým originálem Kunderova Valčíku na rozloučenou a jeho arabským překladem činí 60 %, pro integrační frekvenční zobrazení týchž textů pak činí 49 %. Tato korelace není zrovna oslňující, avšak připouští myšlenku, že změny v délce a frekvenci slov odrážejí význam textu, a navozuje

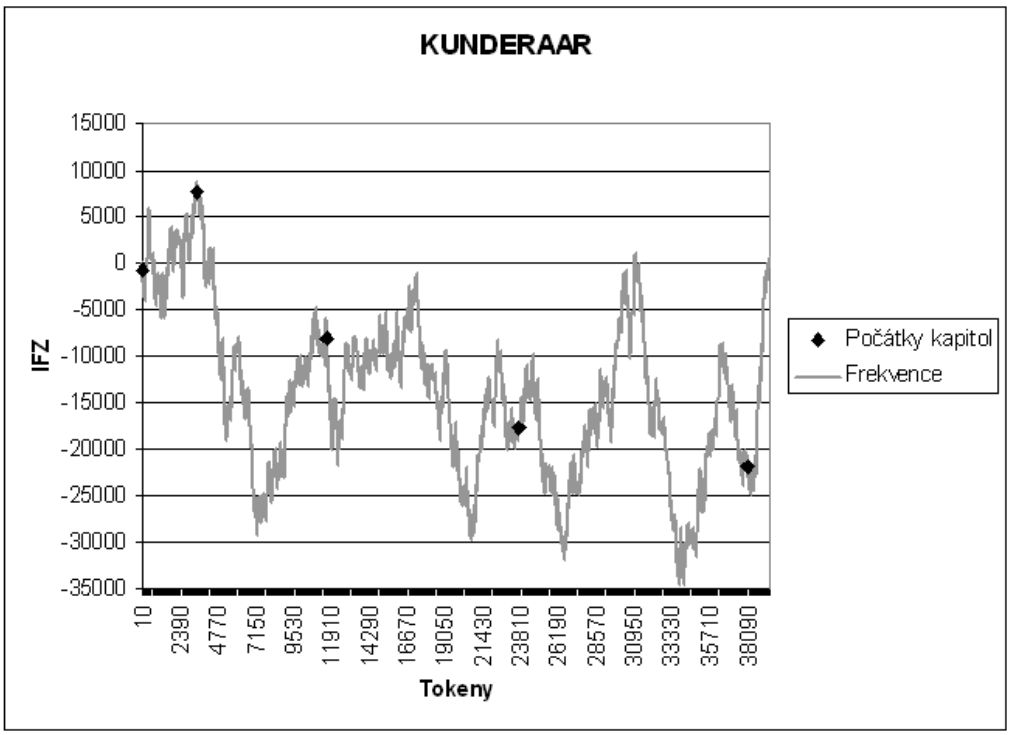
---

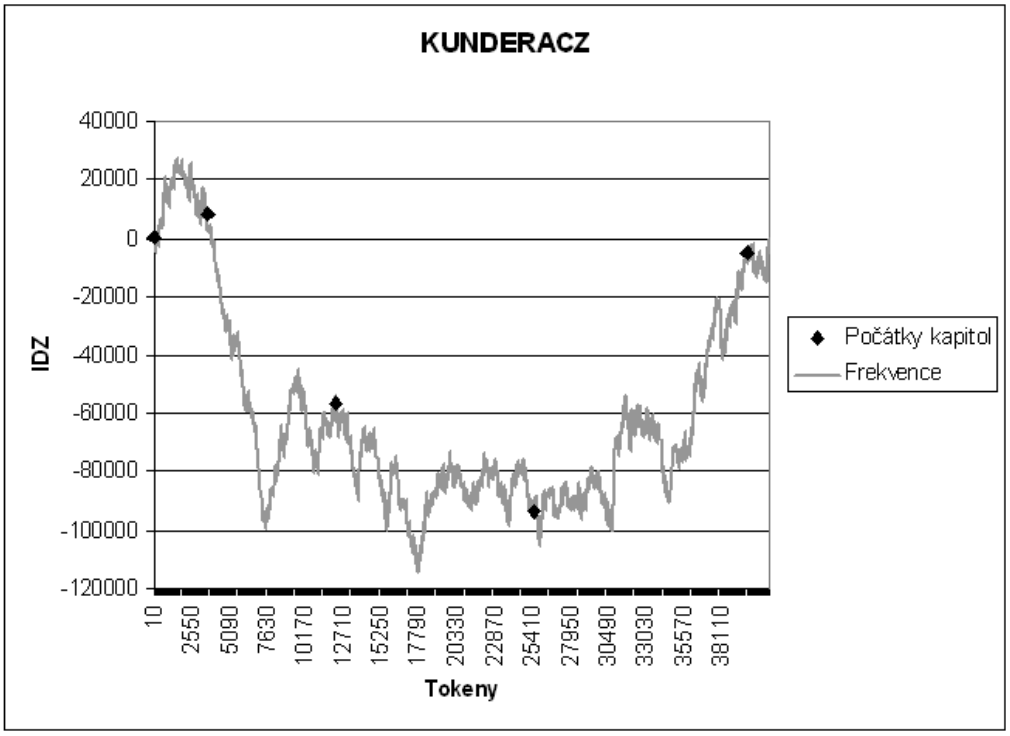
$$h_i = (\frac{im}{n} - im\ div\ n)(g_{im\ div\ n+1} - g_{im\ div\ n}) + g_{im\ div\ n}$$

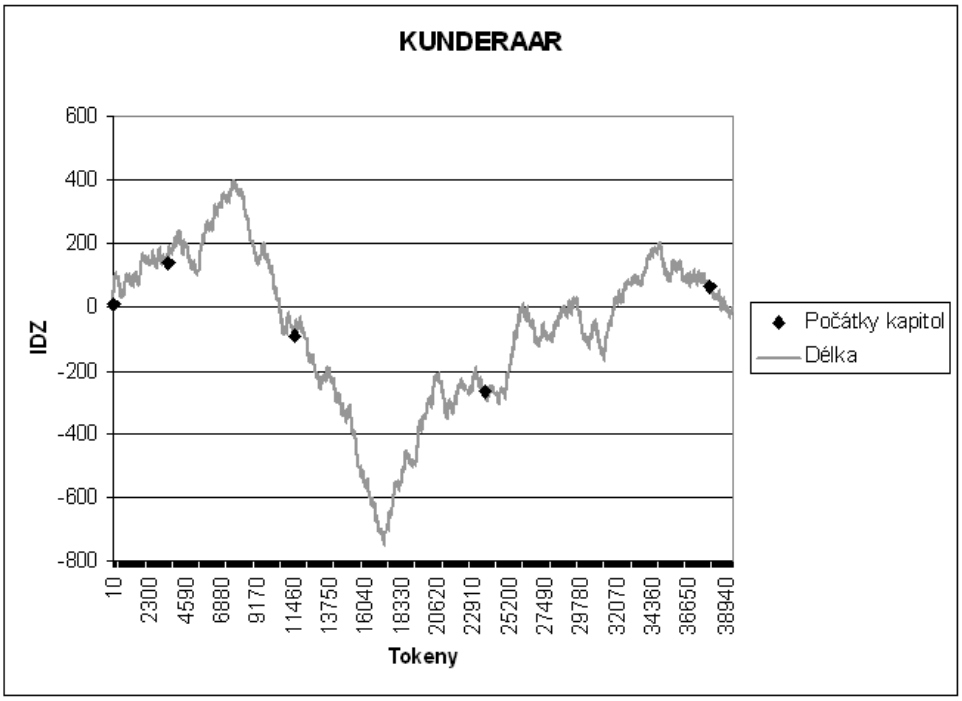
kde div je operand pro celočíselné dělení.

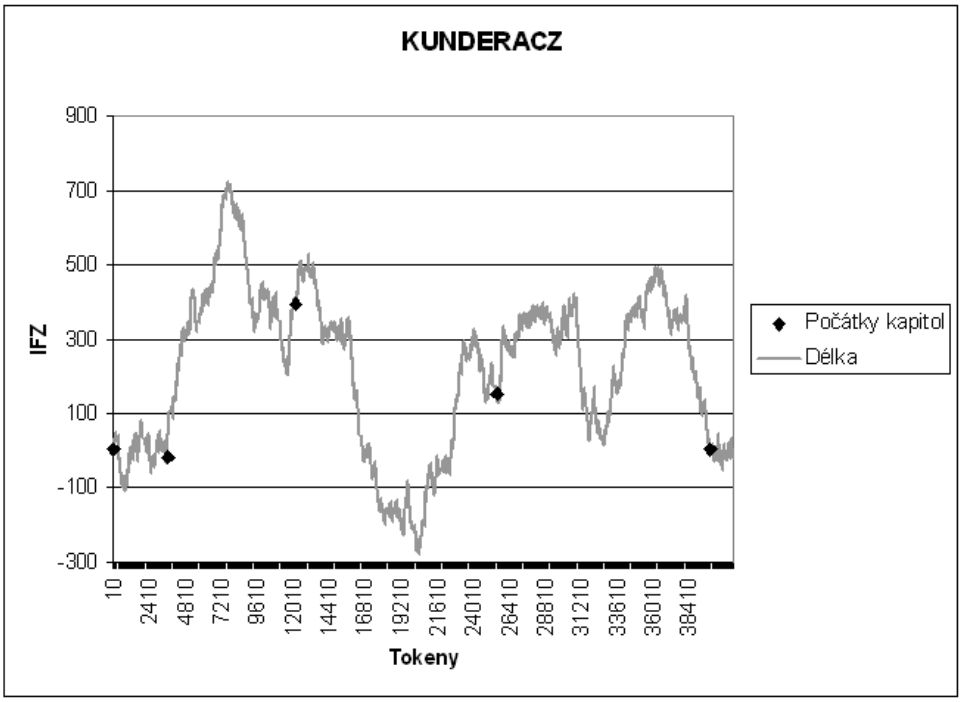
[89] Použity celé korpusy KUNDERAAR a KUNDERACZ

otázku, jestli je možné v grafu vypozorovat nějaké rytmy, které souvisejí například s novou kapitolou, nebo s obsahem textu. Rozdělme si nyní Kunderův Valčík na rozloučenou na jednotlivé kapitoly a podívejme se, jak jejich začátky korespondují se změnami v délkovém a frekvenčním zobrazení:







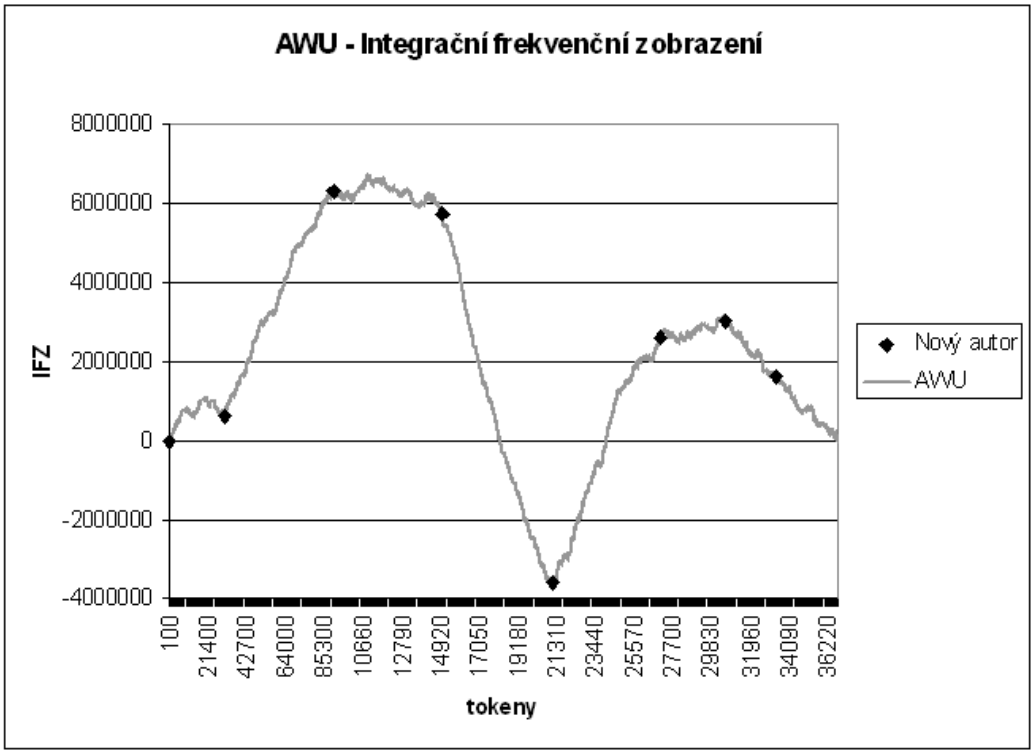


Souvislost mezi změnami směrnice křivek a počátky kapitol není patrná, a to ani u českého originálu, ani u arabského překladu, zjevně tedy souvisí s nějakými méně formálními jevy (obzvláště u Kunderových kapitol nemůžeme mluvit o nějakých sémanticky uzavřených celcích). Některé společné rysy křivek sice můžeme vysvětlit nahlédnutím do textu[90], nicméně této spíše

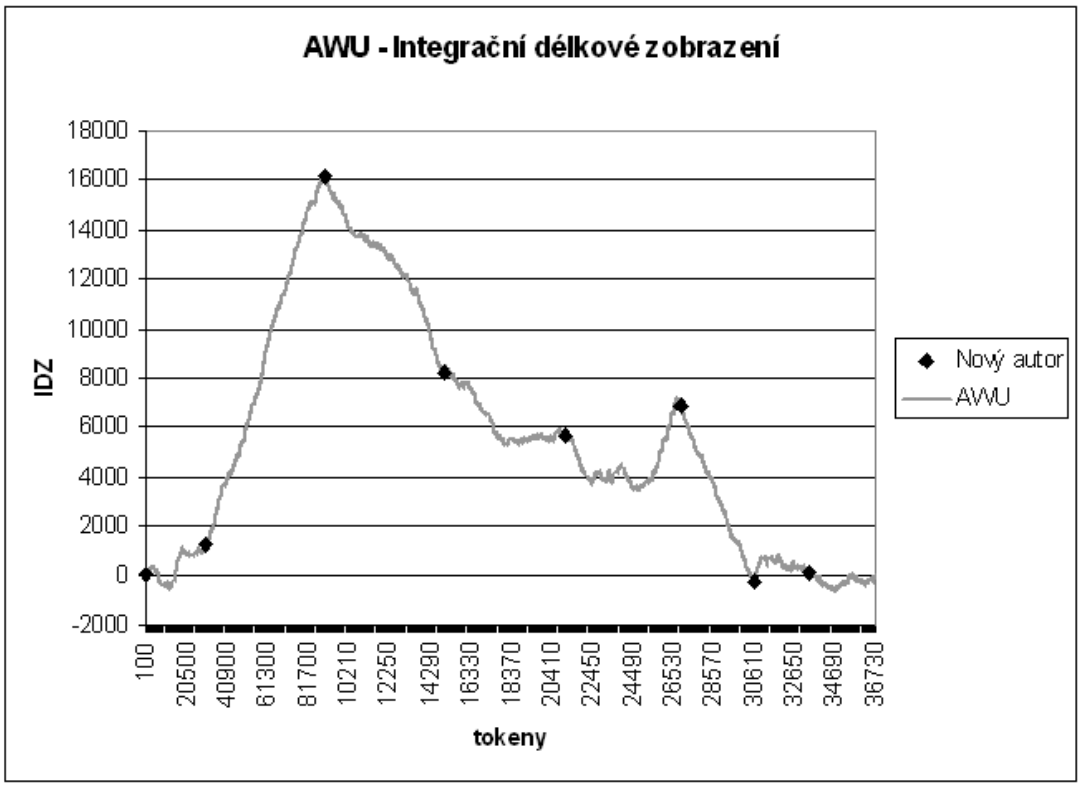
[90] Například zlom okolo 7500. tokenu v češtině a na ekvivalentním místě v arabštině (kolem 7200. tokenu) je dán zřejmě tím, že po pasáži vyprávěné er-formou přichází delší pasáž plná dialogů. To by

literárněvědné analýze můžu ponechat prostor v nějaké příští publikaci. Pro nás je důležitější, že křivky měřené na náhodně zpřeházeném textu vykazují jistou korelaci s křivkami na textu přirozeném (například integrační délkové zobrazení náhodně zpřeházeného JM koreluje 46,6 % s tímtéž zobrazením na korpusu KUNDERAAR). Museli bychom prozkoumat větší množství textů, abychom tento jev mohli zobecnit, nicméně toto zjištění nás inspiruje ke hledání vlastnosti textu, jejíž hodnota bude korelovat pro odpovídající si texty v různých jazycích znatelně více než pro dva nesouvisející korpusy, na rozdíl od těchto dvou popsaných zobrazení – a tím se také budeme zabývat v následující kapitole.

Je ovšem významné, že tyto křivky dokáží odlišit texty napsané různými spisovateli. K posouzení toho se výborně hodí korpus AWU, který je složen z několika novel, povídek a románů různých současných arabských autorů. Vyznačme si na něm začátky literárních celků (každý je od jiného autora):



naznačovalo, že IFZ a IDZ je spíše stylistickým ukazatelem, což je poměrně přesný postřeh, jak si ukážeme později.



To, že délka slova a průměrná četnost jsou pro každého autora charakteristické, není nová informace. Ovšem schopnost těchto zobrazení pomoci opticky určit změnu autora[91] nás přímo vybízí k vytvoření algoritmu na rozpoznávání autorství, který by na nich byl založen. Tyto grafy jsem uvedl spíše jako ilustraci toho, že naše teoretické úvahy nad strukturou nadvětných celků v textu mohou mít i praktické uplatnění.

[91] Každou změnu autora v tomto korpusu doprovází prudká změna směrnice v alespoň jednom z uvedených grafů. Například Sulaymān Kāmil (druhý autor) na rozdíl od svého předchůdce Muḥammada abū Maᶜtūqa používá delší a častější slova. Anīs Ḥūrī (3. v pořadí) má stejně dlouhá slova jako Muḥammad Ḥusayn, který následuje, ten ovšem používá mnohem vzácnější slova.

5.4 Shrnutí

Závislost frekvence slova na frekvenci slov v jeho okolí není přesvědčivá, naproti tomu vztah mezi délkou slova a průměrnou délkou slov v sousedství se jeví být charakteristickou pro každý jazyk. Ovšem křivka této závislosti není snadno popsatelná. Daleko lépe můžeme charakterizovat vztah mezi délkou věty a průměrnou délkou vět v jejím sousedství jako přímou úměru, podobně jako průměrnou délku slova ve větě a průměrnou délku slova ve větách v sousedství této věty – jedná se *cum grano salis*[92] o lineární závislosti, což by nás mohlo přimět k uvažování o tom, že se v textu utvářejí shluky dlouhých a krátkých vět a delší celky, které obsahují průměrně dlouhá, nebo naopak krátká slova.

V nadpisu jsem sliboval pohled do nadvětných struktur a myslím, že po této kapitole můžeme definitivně zapomenout na představy generativistů, že věta tvoří nezávislou jednotku jazyka.

Ve třetí podkapitole otvíráme otázku, jestli je toto střídání pasáží s dlouhými a krátkými slovy (četnějšími a vzácnějšími slovy) nějak podmíněno sémanticky. Pro lepší porovnávání definujeme a využíváme integrační délkové zobrazení (IDZ) a integrační frekvenční zobrazení (IFZ). Ukazuje se však, že korelace datových řad naměřených pro různé jazykové mutace týchž textů se příliš neliší od korelací pro náhodné texty – daleko slibněji vypadá zjištění, že IDZ a IFZ nám může posloužit jako stylometrický indikátor, ať už v literární vědě, nebo při určování autorství textu, a pokud se mi (nebo komukoli jinému) při dalším zkoumání podaří rozvinout tuto jeho vlastnost do funkčního algoritmu, stane se jednou z prakticky uplatnitelných myšlenek této studie.

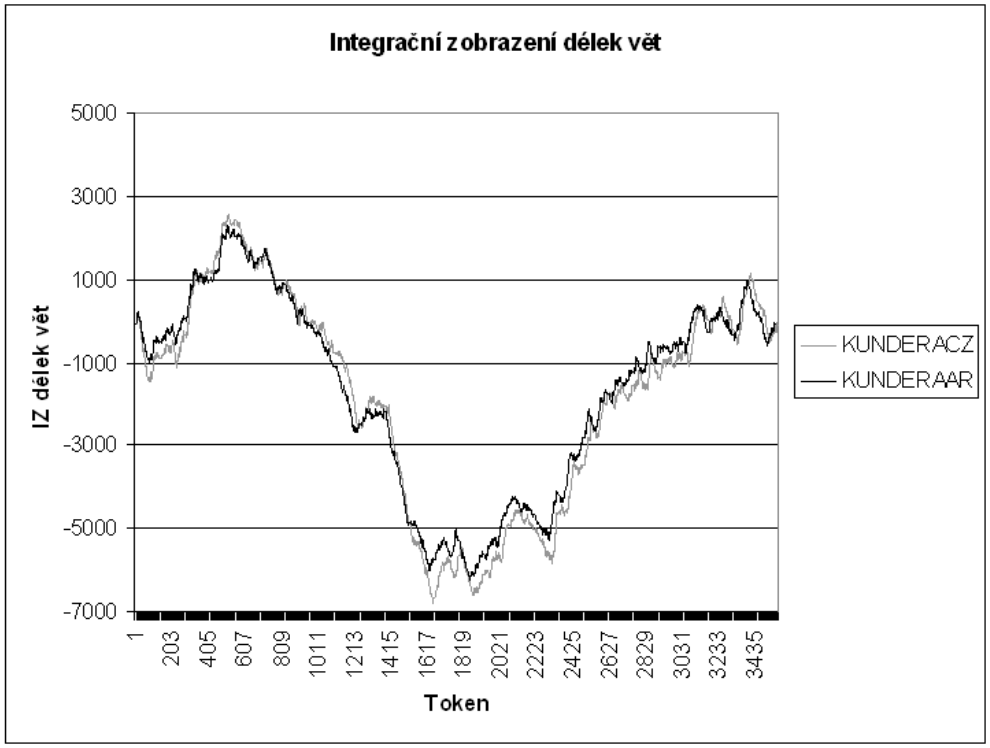
Původní představy se ovšem nenaplnily, a tak jsem se zdržel formulace předpokládaných hypotéz. Hledání zobrazení, které dokáže zachytit význam textu nezávisle na jeho jazyce, se však nevzdáváme a budeme v něm pokračovat v následující kapitole.

[92] Musíme mít na paměti, že lineární vztahy, které odkrýváme, se mohou jevit jako lineární pouze v onom krátkém intervalu, kde jsou měřitelné. Ovšem pokud uznáme, že nemá příliš smysl hovořit o slovech o délce například 100 znaků, nemusí nás z praktického hlediska zajímat, že pro tuto hodnotu by se naše přímka změnila v hyperbolu, nebo nějakou jinou, hůře identifikovatelnou křivku, jiná situace ovšem nastane, pokud s tímto vztahem budeme chtít dále pracovat a odvodit z něj vztah jiný – je dobré si uvědomit, v jakém intervalu platí.

6. Kombinatorické zobrazení

V předchozí kapitole jsme se dozvěděli, že je sice přirozené, aby překladatel pasáže s relativně dlouhými nebo méně četnými slovy překládal jako pasáže s relativně dlouhými nebo méně četnými slovy a naopak, že to ale není nic nevyhnutelného, že střídání takových pasáží je spíše spjato se stylem než se smyslem textu (podle naší definice jazyka je styl jakási *varianta metody komunikace*).

Ptejme se, jakým způsobem je třeba zobrazit text, aby byl výsledek jeho zobrazení stejný pro originál i pro překlad. Dobré výsledky přináší například integrační zobrazení délek vět.

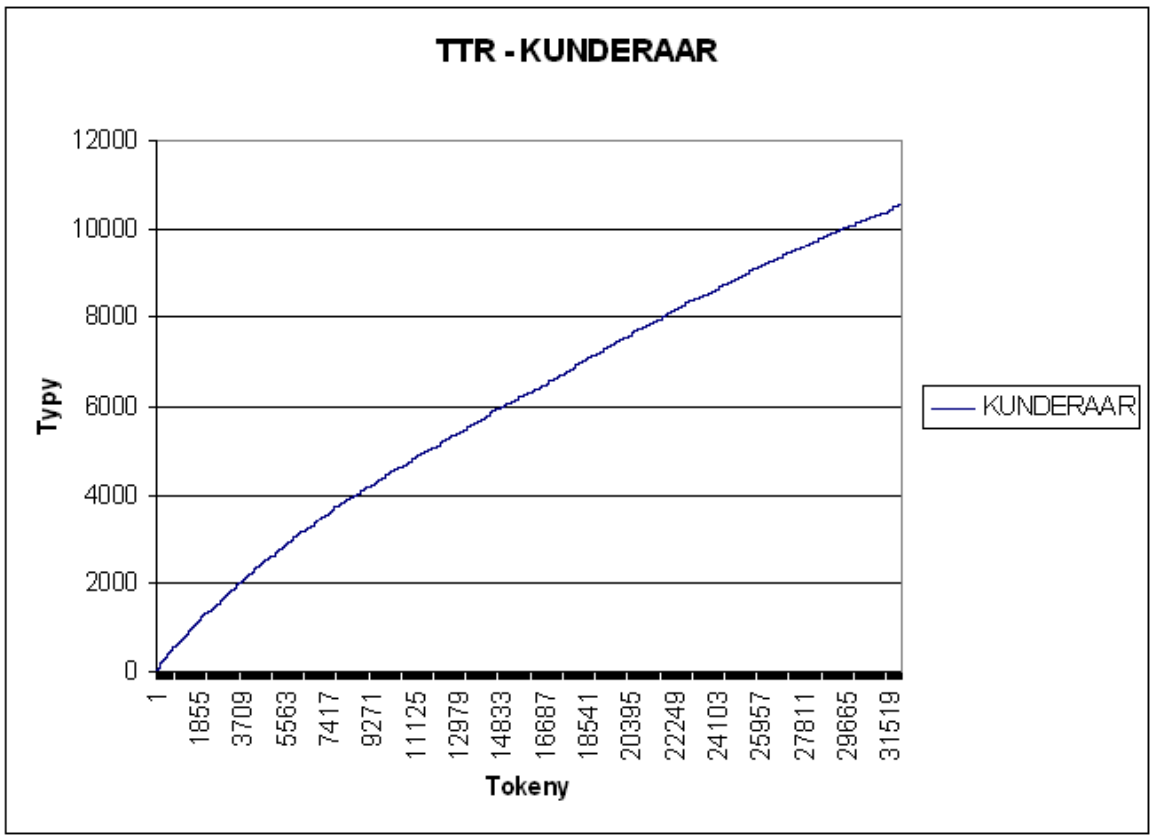


Grafy si opticky pěkně odpovídají a korelace mezi datovými řadami pro český a arabský text dosahuje 99,1 %. Jenomže věta je značně formální prvek a pokud jeden překladatel kopíruje hranice vět původního textu, ještě to neznamená, že totéž musejí dělat všichni. Následující tvrzení nemám ověřené, ale zdá se mi, že čeští překladatelé arabských textů daleko více přizpůsobují hranice arabských vět českým zvyklostem a ortografii, než překladatelé arabští. Hledejme zobrazení, které by osvětlovalo struktury, kterých si překladatel při své práci vůbec není vědom a jejichž kopírování je *nutností, kterou je obtížné obejít*.

Když mluvčí začíná nový významový celek, znamená to často, že mluví o věcech, o kterých ještě v textu nemluvil. K tomu chtě nechtě používá typy, které ještě v textu nepoužil. Respektive používá jich více, než na konci významového celku a totéž musí udělat i překladatel. Tohoto

předpokladu využijeme v tzv. **kombinatorickém zobrazení textu**. To vychází ze vztahu tokenů k počtu použitých typů[93], v následující pasáži si ho představíme blíže.

Viděli jsme, že i poměrně jednoduchá měření přinášejí zajímavé nové výsledky a že není příliš třeba pouštět se do vysokých abstrakcí. V tomto případě se však bez ní neobejdeme, neboť vztah typů a tokenů má pro všechny texty stejný charakter (a to dokonce i pro „texty" s přeházenými tokeny), který je dán rozdělením slov podle Zipfovské distribuce[94]. Vypadají přibližně takto:



Na první pohled patrný tvar křivky vzniká samovolně podle stochastických pravidel, význam textu se odráží v drobných odchylkách a vlnách, které jsou téměř nezřetelné. Jak tyto drobné odchylky zviditelnit? Vytvoříme všechny permutace tokenů v daném textu. Na těchto permutacích změříme TTR. Všechny naměřené křivky zprůměrujeme a výslednici odečteme od křivky naměřené na zkoumaném textu.

---

[93] Obvyklý název tohoto vztahu v západní literatuře je *type-token ratio*, nebo *type-token relation*, zkráceně TTR. Procházíme text po jednotlivých tokenech a zapisujeme si do slovníku každý nový typ, který se v textu dosud nevyskytoval. Každému tokenu zároveň přiřadíme počet typů v tomto slovníku.

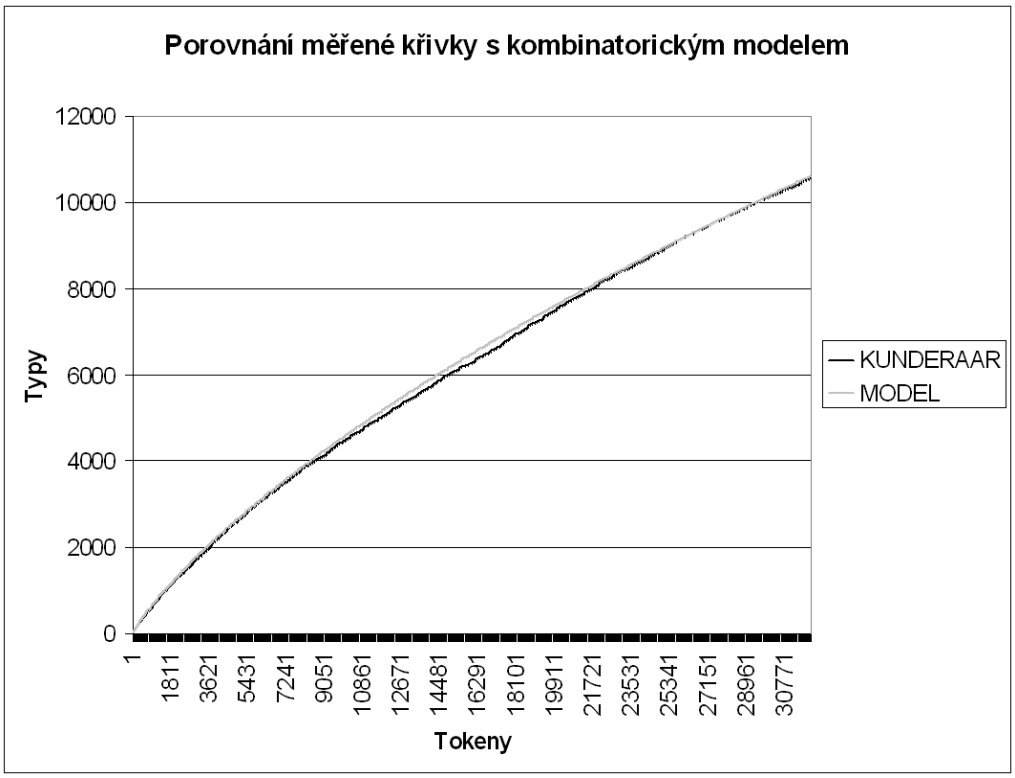
[94] Jak jsem ukázal v článku (Milička 2009). Z tohoto článku budeme vycházet i při dalších úvahách v této kapitole.

Ano, tento postup je komputačně neproveditelný, ovšem permutování tokenů v textu můžeme nahradit modelem, který přesně nahrazuje výslednici zprůměrovaných TTR následujícím vzorcem[95]:

$$V = \sum_{i=1}^{M} \left[1 - \frac{(d-N)!(d-f_i)!)}{d!(d-N-f_i)!} \right] \qquad \text{pro } d \geq N + f_i \wedge N; f_i; d \in N$$

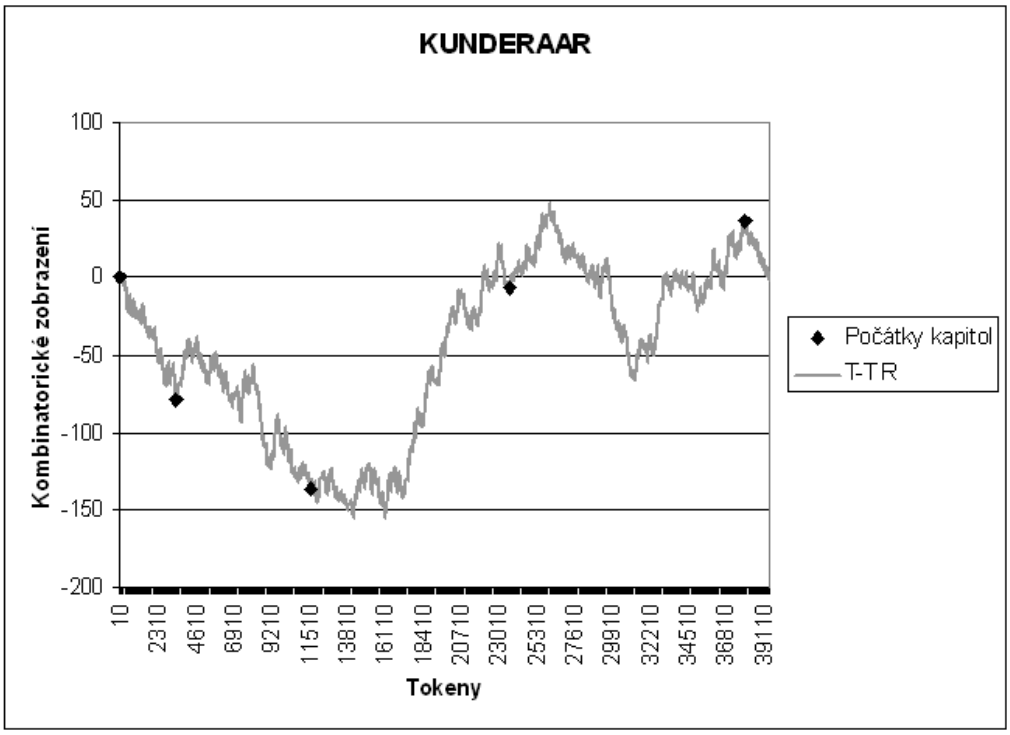
kde V je počet typů, N počet tokenů, f_i frekvence typu, který je $i$tý v pořadí a d celkový počet různých typů v textu.

Díky tomuto modelu se dostaneme ke křivce, která se liší od oné naměřené jen velmi málo:



Ovšem ony malé rozdíly (skutečně malé, R^2 = 99,97 %) jsou významné, pokud jde o záměry autora a význam textu – odečteme křivku kombinatorického modelu od naměřené křivky:

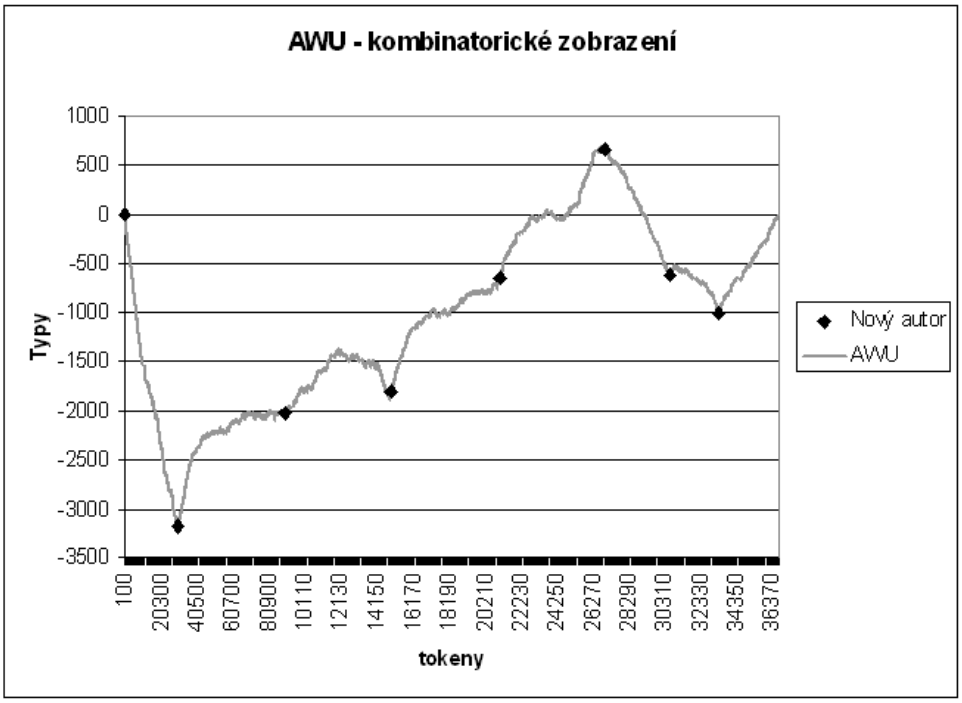
[95] Kombinatorický model TTR, odvození je provedeno ve výše zmiňovaném článku (Milička 2009), do vztahu vstupují frekvence všech typů.



První a poslední člen kombinatorického zobrazení je roven nule. V místech, kde je větší směrnice, jsou slova, která do té doby nebyla v textu použita více, než by odpovídalo statistickým zákonitostem. Jednoduše řečeno, tam, kde křivka roste, začíná nový významový celek, tam kde klesá, autor opakuje stará témata.

Ani tady není příliš zřetelné, že by kapitoly začínaly v místech, kde graf roste. Když se ale podíváme na zlomové body blíže, zjistíme, že skutečně souhlasí nikoli s formální, ale významovou stránkou textu. Například křivka začne prudce stoupat u 18310. tokenu, kde začíná vyprávění o lovu psů, což je v knize naprosto ojedinělé téma. Po začátku 4. kapitoly stoupá křivka, neboť na scénu přichází nová postava, Klímova manželka.

Existuje jistá pozitivní korelace mezi kombinatorickým zobrazením a integračním délkovým zobrazením – u KUNDERAR činí 65 %, u KUNDERACZ 24 %, ovšem u korpusu AWU se souvislost nepotvrdila – korelace je -39 %. Podobně jako integrační zobrazení, i kombinatorické nás dokáže upozornit na změnu autora:

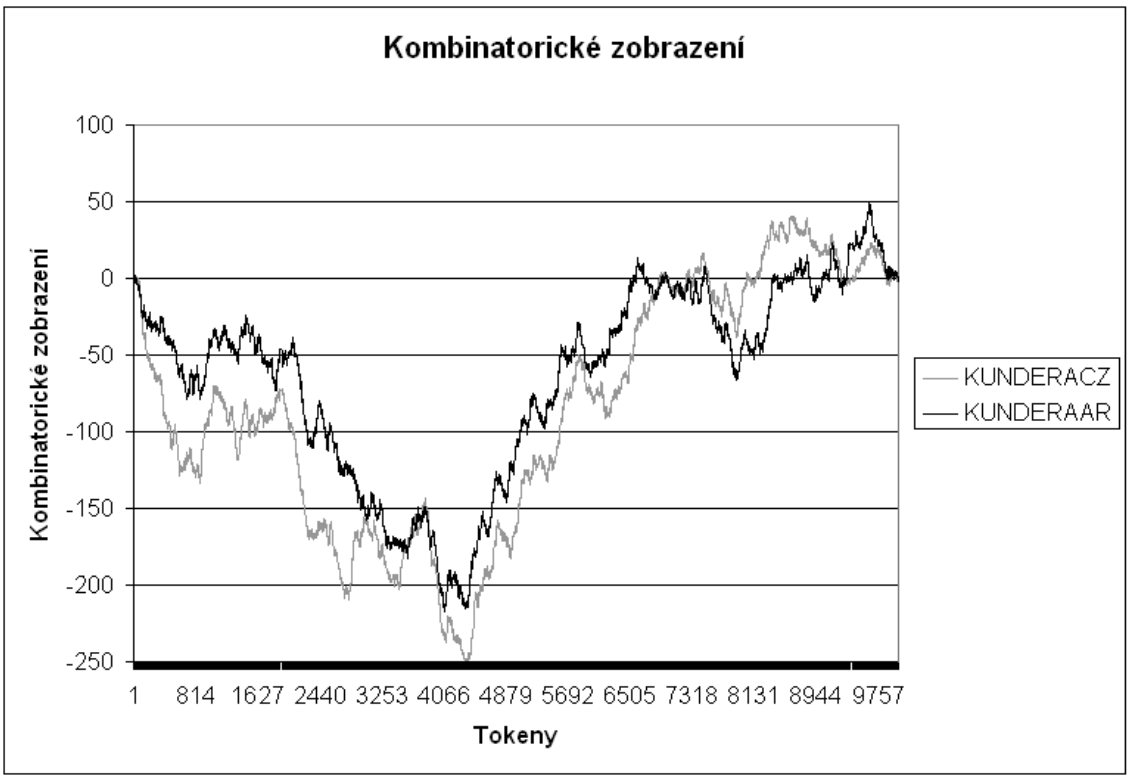


Se změnou autora – a zároveň začátkem rozsáhlého významového celku – začne křivka prudce stoupat a tvoří tak pro každý celek jakousi vlnu.

Ovšem hlavně nám toto zobrazení dovoluje formulovat následující hypotézu:

Hypotéza č. 18: **Pokud je text A překladem textu B, tehdy a jen tehdy kombinatorické zobrazení textu A významně koreluje ke kombinatorickému zobrazení textu B.**

Tato hypotéza platí na našich korpusech KUNDERAR a KUNDERACZ, kde Pearsonův korelační index činí 93 %. Že jsou si ty dvě datové řady podobné je znát i na grafu:



Naopak s datovou řadou změřenou na korpusu s náhodně zpřeházenými tokeny je korelační koeficient jak pro český text tak pro jeho arabský překlad menší než 0. V dalším výzkumu je potřeba rozvést, co všechno je možné považovat za významnou korelaci. A ovšem – ne zrovna přesné je také vyjádření, že text A je překladem texu B, mezi překladem, převyprávěním a naprostou významovou nezávislostí je mnoho odstínů.

Ukazuje se, že tato hypotéza platí i pro jiné dvojice jazyků (např. čeština a angličtina) a jiné typy textů, ovšem k opravdové koroboraci by bylo třeba vyzkoušet ji na reprezentativním vzorku textů. Pokud by se osvědčila, znamenalo by to, že prostřednictvím kombinatorického zobrazení skutečně nahlížíme do významu textu a nikoli jen do jeho formální struktury.

7. Závěr

V první kapitole této studie jsem stanovil demarkační kritérium pro vědecké hypotézy, kterému jsem přizpůsobil svou metodiku a kterým jsem se řídil i v průběhu celé práce.

Za vědeckou hypotézu považujeme každé intersubjektivně vyvratitelné non-existenciální tvrzení a každé intersubjektivně ověřitelné existenciální tvrzení.

Pokud čtenář s tímto kritériem nesouhlasí, nemusí zavrhovat celou studii, neboť hypotézy, které navrhuji, odpovídají také epistemologii popperovské.

Co vlastně říká oněch 18 falzifikovatelných obecných výroků (vyvratitelných non-existenciálních tvrzení), okolo nichž je tato práce vystavěna? Pojďme se podívat na nejdůležitější z nich:

Druhá kapitola rozebírá frekvenci slov. Ukazuje se v ní, že rozložení častých a vzácných slov ve větě není náhodné. Například ve všech zkoumaných korpusech se na koncích vět soustředila poměrně vzácná slova. Podobné tendence bylo možno vypozorovat na různých pozicích ve větě, takže v průměru vykazovaly věty určitou strukturu. Tato struktura byla podobná pro věty o různém počtu tokenů a byla typická pro každý jazyk. Zároveň ale bylo možné najít i podobnosti pro různé jazyky, ať už příbuzné, nebo nepříbuzné. V souvislosti s arabštinou nás musí zaujmout, že struktury naměřené na korpusu v moderní spisovné arabštině a na korpusech středověkých arabských autorů se značně podobají.

Podobně třetí kapitola pojednává o délce slov. Také různě dlouhá slova se ve větě nedistribuují náhodně a nacházíme zde určité struktury, přičemž tyto struktury občas vykazují vysokou zápornou korelaci s těmi, které jsme nalezli v předchozí kapitole. Také struktura délek v průměrné větě je typická pro každý jazyk a je podobná pro věty o různém počtu tokenů. Tyto struktury vykazují korelaci mezi nepříbuznými jazyky v daleko větší míře než v předchozím případě a ukazuje se, že tak dokážeme dobře rozlišit přirozený text od „textu“ s náhodně transponovanými tokeny.

Tyto dvě kapitoly naznačují, že co platí pro frekvenci, může platit v nějaké podobě i pro délku. Pokud najdeme nějaké praktické uplatnění pro jakýkoli vztah, který se týká frekvence, může být efektivnější ji nahradit délkou, neboť se dá rychleji změřit.

Kapitola 4 tvoří jakési *intermezzo* – pokouší se dát do souvislosti délku věty a průměrnou frekvenci slov v ní, přičemž naráží na Menzerathův-Altmannův zákon. Společně se vztahem délky a frekvence slov ve větě tvoří jakési trio regresí, které je možné algebraicky odvozovat jednu z druhé za použití parametrů ostatních dvou regresí, což je v zápětí názorně ukázáno.

Podobnou metodou by bylo možno popsat více vztahů – při řešení této kapitoly jsem se nemohl zbavit pocitu, že kdybychom začali čistě mechanicky, ať už náhodně, nebo systematicky, dávat do vztahů různé veličiny, které vůbec lze kvantifikovat a měřit na textu, zjistili bychom, že v jazyce všechno souvisí se vším a jen málo vlastností textu nemá vůbec žádný vzájemný vztah.

Znalostí z této kapitoly využívá kapitola následující, ve které se ovšem ukazuje, že ne vždy jsou vztahy, které nalezneme pro délku, zřetelné i pro frekvenci. Nepodařilo se například stanovit metodu falzifikace pro hypotézu, že na frekvenci slova závisí průměrná frekvence slov v jeho sousedství, neboť data byla příliš chaotická. Naproti tomu pro vztah délek tokenů a průměrných délek tokenů v jejich sousedství byl dobře rozpoznatelný a i když bude zřejmě těžké popsat jej nějakou rovnicí, může nás uspokojit zjištění, že je pro texty napsané v různých jazycích velmi podobný. Podobným způsobem nalezneme vztah na nadvětné úrovni – průměrná délka vět v pravém i levém bezprostředním okolí vět, které jsou dlouhé n písmen, je přímo úměrná n. Tato hypotéza byla vyvrácena pro dva francouzské korpusy, pro ostatní testované jazyky však platí. Tři arabské korpusy, které máme k dispozici, jsem využil k tomu, abych tuto hypotézu zúžil na arabský jazyk. Další hypotéza – Průměrná délka slov ve větách v pravém i levém bezprostředním okolí vět, které obsahují slova o průměrné délce n písmen, je přímo úměrná n – platí nejen pro věty, ale i pro mechanicky rozdělené celky o určitém počtu písmen. Zřejmě tedy v textu najdeme delší pole s relativně krátkými slovy a delší pole s relativně dlouhými slovy. Tuto hypotézu jsem zkusil využít a definoval jsem integrační zobrazení. Následně se ukázalo, že integrační frekvenční zobrazení textu a integrační délkové zobrazení textu dobře reflektuje stylové změny, avšak pohled do nadvětných struktur nabízí jen mlhavý.

Poslední, šestá kapitola je zasvěcena zkoušení, jestli by lepší pohled nemohlo nabídnout kombinatorické zobrazení textu, které využívá vztah typů a tokenů. Skutečně se ukazuje, že toto zobrazení reflektuje změny na významové a ne na stylistické rovině a že může platit poslední hypotéza: pokud je text A překladem textu B, tehdy a jen tehdy kombinatorické zobrazení textu A významně koreluje ke kombinatorickému zobrazení textu B. Tak máme k dispozici metodu, která zobrazuje význam textu nezávisle na jazyce, ve kterém je napsaný.

Je možné, že jednotlivé hypotézy nebo algoritmy přinesou někdy praktický užitek, nicméně hlavní přínos této studie vidím v otevření nových možností pro teoretické úvahy o jazyce. Nacházíme další pravidelnosti textu, o nichž nemá mluvčí tušení, a ukazuje se, že tyto struktury jsou vzájemně propojeny. Někdy jsou charakteristické pro určitý jazyk, jindy jsou typické pro jakýkoli přirozený text. Naopak některé vnější aspekty textu jsou nezávislé na jazyce, v jakém je napsán, a jsou pevně dány jeho významem. Povzbuzující je, že tyto struktury můžeme zobrazovat

pomocí velmi jednoduchých algoritmů, pomocí složitějších úkonů, nebo na vyšším stupni abstrakce možná dosáhneme ještě zajímavějších výsledků.

Tato studie nevznikla v knihovně, ale v reálném kontaktu se zkoumaným jazykem a textem, zejména arabským. A také s lidmi, kterým vděčím za inspiraci, ať už to byli lingvisté, filologové, ekonomové, nebo neměli s vědou nic společného. Myšlenky se ovšem rodily také při procházkách, nebo v houpací síti a neučesané přicházely ve snu. Často ležely dlouhou dobu rozpracované v šuplíku a teprve termín odevzdání diplomové práce mě přiměl dát jim podobu, které by porozuměli ostatní. Forma této studie tomu odpovídá. Vynaložil jsem značné úsilí, abych tyto myšlenky sjednotil a poskládal tak, aby dávaly dohromady smysl, avšak přesto se nemohu zbavit dojmu, že se skutečně jedná jen o jakýsi sled „vhledů do struktury arabského textu" (a nutno dodat, že nejen arabského), a nezbývá mi, než doufat, že další soustavnější práce na tomto poli nabídne systematizaci těchto jevů, nebo je včlení do širšího rámce.

8. Literatura

Abee, Susan (2000). Word Length Distribution in Arabic Letters. *Journal of Quantitative Linguistics*, Volume 7, Issue 2 August 2000, s. 121–127. ISSN: 0929-6174.

Alnajem, Salah (2005). A computational approach to the variations in Arabic verbal orthography. Computer Speech and Language 19 (2005). Elsevier, s. 275–299. ISSN: 0885-2308.

Altmann, Gabriel – Köhler, Reinhard – Vulanović, Relja (2006). Laws in Quantitatie Linguistics. Dostupné z: http://lql.uni-trier.de/ [13. 8. 2010]

Bahbouh, Charif – Bahbouhová, Martina (2001). 365+1 arabské přísloví a mudrosloví. 1. vydání. Praha: Dar ibn rushd. 142 s. ISBN: 8086149-29-3.

Beesley, Kenneth (1996). Arabic finite-state morphological analysis and generation. In *COLING-96: Papers Presented to the 16th International Conference on Computational Linguistics*, s. 89–94. Dostupné z: http://acl.ldc.upenn.edu/C/C96/C96-1017.pdf [13. 8. 2010].

Cvrček, Václav (2009). Regulace jazyka a koncept minimální intervence. 1. vydání. Praha: Nakladatelství Lidových novin. 123 s. ISBN: 978-80-7106-600-2.

Kemlík, Vítězslav (2010). *Climate gate česky*. Dostupné z url http://kremlik.blog.idnes.cz/ [13. 8. 2010].

Kiraz, George (2001). *Computational Non-linear Morphology. With Emphasis on Semitic Languages*. 1. vydání. Cambridge: Cambridge University Press. xxi, 171 s. ISBN: 0-521-63196-3.

Kirchhoff, Katrin a kol (2006). Morphology-based language modeling for conversational Arabic speech recognition. Computer Speech and Language 20 (2006). Elsevier, s. 589–608. ISSN: 0885-2308.

Kropáček, Luboš (1996). *Islámský fundamentalismus*. 1. vydání. Praha: Vyšehrad. 263 s. ISBN: 80-7021-168-7.

Mair, Christian (2006). Tracking Ongoing Grammatical Change and Recent Diversification in Present-day Standard English: The Complementary Role of Small and Large Corpora. In: *The Changing Face of Corpus Linguistics* VII. Amsterdam: Rodopi. s. 355-376. ISBN 90-420-1738-4.

Milička, Jiří (2009). Type-token & Hapax-token Relation: A Combinatorial Model. *Glottotheory 2/1*. Trnava, s. 99-110. ISSN: 1337-7892.

Popper, sir Karl Raymund (1997). *Logika vědeckého zkoumání*. Přel. Jiří Fiala. 1. vydání. Praha: Oikomenh. 617 s. ISBN: 80-86005-45-3.

Prokosch, Eduard (1933). *Reviewed work: Selected Studies of the Principle of Relative Frequency in Language* by George Kingsley Zipf. *Language*, Vol. 9, No. 1 (Mar., 1933). Washington D. C., s. 89-92. ISSN: 00978507.

Smrž, Otakar (2007). ElixirFM: implementation of functional Arabic morphology. In ACL 2007 Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, s. 1–8, Prague, Czech Republic, Dostupné z: http://www.aclweb.org/anthology/W/W07/W07-0801 [13. 8. 2010].

Strauss , Udo – Altmann, Gabriel (2006). Word length and frequency. In Laws in Quantitatie Linguistics. Dostupné z: http://lql.uni-trier.de/ [13. 8. 2010].

Volín, Jan (2007). Statistické metody ve fonetickém výzkumu. 1. vydání. Praha: Nakladatelství Epocha. 344 s. ISBN: 978-80-87027-54-7.

Wolfram, Stephen (2002). *A New Kind of Science*. Champaign: Wolfram Media. 1197 s. ISBN: 1-57955-008-8.

Wyllys, Ronald (1981). Empirical and Theoretical Bases of Zipf's Law. *Library Trends* 30(1). s. l., s. 53-64. Dostupné z: http://www.ischool.utexas.edu/~wyllys/EmpiricalAndTheoretical.pdf [13. 8. 2010].

Zipf, George Kingsley (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, Massachusets: Harvard University Press, 1932.

9. Charakteristika použitých korpusů

KOMPLET Soubor veršů předislámských básníků, zejména z básní sbírky *Muᶜallaqāt*. Většinu mají básníci *Zuhayr ibn abī Salmā* a *Imru ´l-Qays* se svými básněmi psanými v metru ṭawīl, avšak jsou obsažena i další metra (wāfir, ramal) a další básníci. Cca 9 000 slov (extrémně malý rozsah).

AWU Soubor románů, povídek a novel psaných moderní spisovnou arabštinou od mnoha různých autorů (například *Sulaymān Kāmil*, *Fā´iq Muḥammad Ḥusayn*, *Nadya Ḥūst*...) organizovaných v *Arabic Writers´ Union*. Cca 400 000 slov. Není sice plně homogenní[96], to se však při testech ukázalo jako nepodstatné.

AWLAD Část slavného románu *Nağība Maḥfūẓe* (Poválečného káhirského spisovatele) Děti naší čtvrti (*Awlād ḥāratinā*). Cca 21 000 slov, pro svůj malý rozsah se hodí spíše pro ověřování funkčnosti algoritmů.

COOPER Soubor 3 nejslavnějších románů od *Jamese Fenimora Coopera – The Deerslayer*, *The Last of the Mohicans*, *The Prairie* . Tento korpus se dobře osvědčil pro měření různých vlastností textu, neboť zahrnuje jednoduché lineární vyprávění jednotným stylem bez rušivých prvků. Cca 500 000 slov.

MOHICAN Samotný román *The Last of the Mohicans* od *Jamese Fenimora Coopera.* Cca 150 000 slov.

MLOCI Soubor tří větších děl od *Karla Čapka – Války s Moloky*, *Krakatit*, *Hovory s TGM* a několika jeho menších publikací (jako například *Anglické listy*). Cca 400 000 slov.

JM Korpus zahrnuje Knihu zvířat (*Kitāb al-ḥayawān*) od *al-Ğāḥiẓe* (slavného arabského polyhistora z 9. století) a slavnou geografickou publikaci Rýžoviště zlata a doly drahokamů (*Murūğ ad-ḏahab wa-´lmaᶜādin al-ğawāhir*) (kniha byla také Ivanem Hrbkem přeložena do češtiny) od al-Masᶜūdīho z 10. století. Čítá asi 800 000 slov, byl zbaven poetické složky a rejstříků.

---

[96] Homogenita vzorku nám zjednodušuje interpretaci měřených dat. Jan Volín varuje před zahrnováním zástupce různých populací do jednoho vzorku, (str. 87, Volín 2007). Například rozložení průměrné výšky hlasu se řídí normálním rozdělením (Gaussovým), ovšem pokud zahrneme do stejného vzorku muže i ženy, získáme rozložení se dvěma vrcholy, na kterém nebudeme moci použít statistické nástroje vytvořené pro normální rozdělení.

CHALDUN Celé dějepisné dílo *ᶜAbdarraḥmāna ibn Ḥaldūna* ze 14. století vytváří korpus čítající celkem asi 1 400 000 slov.

ZOLA Korpus je tvořen několika romány *Émila Zoly* (*Nana, La Fortune des Rougon, Le Ventre de Paris...*). Cca 1 000 000 slov.

HUGO Korpus je tvořen několika romány *Viktora Huga* (*Les Misérables...*) a *Alexandra Dumase* (např. *Le comte de Monte-Cristo*). Celkem má přibližně 1 500 000 slov.

KORANAR Arabský originál Koránu. Je zarovnán jako paralelní korpus se svým českým překladem. Nutno poznamenat, že jazyk a styl Koránu je natolik specifický, že jakékoli závěry z tohoto korpusu je nebezpečné zobecňovat. Byl používán v době, kdy nebyl k dispozici korpus KUNDERAAR pro testy, které vyžadovaly paralelní korpus. Cca 80 000 slov.

KORANCZ Český překlad Koránu. Cca 130 000 slov.

KUNDERAAR Román *Valčík na rozloučenou* od *Milana Kundery*. Je zarovnán jako paralelní korpus se svým arabským překladem. Ideální narativní text psaný jednotným jazykem a stylem. Necelých 40 000 slov.

KUNDERACZ Arabský překlad (překladatel neznámý, překlad šel zřejmě přes francouzštinu) románu Valčík na rozloučenou (*Fāls al-widāᶜ*) od *Milana Kundery*. Je zarovnán jako paralelní korpus se svým českým originálem. Cca 40 000 slov.

Příloha A

Délkové a frekvenční zobrazení v praxi

Abych nerušil výklad, rozhodl jsem se dát příklad měření z kapitoly 5 do zvláštní přílohy, na kterou v této kapitole odkazuji. Následující tabulka je výňatkem z korpusu AWU a jeho délkového a frekvenčního zobrazení:

	Délka	Levý	Pravý	Četnost	Levý	Pravý
أرغب	4	4	2	20	95	10425
في	2	4	5	10425	20	5
تعليق	5	2	2	5	10425	1087
كل	2	5	3	1087	5	437
شيء	3	2	2	437	1087	1087
كل	2	3	3	1087	437	437
شيء	3	2	3	437	1087	1123
حتى	3	3	3	1123	437	78
ولو	3	3	5	78	1123	2
ترنحت	5	3	7	2	78	43
الجدران	7	5	2	43	2	7811
من	2	7	4	7811	43	20
وطأة	4	2	8	20	7811	5
المسامير	8	4	7	5	20	1
والثياب	7	8	9	1	5	5
وذكريات	9	7	8	5	1	1
المعلقة	8	9	6	1	5	171

Na těchto datech si ilustrujeme použitá měření – nejprve z kapitoly 5.1. V uvedeném textu mají například slova s pěti výskyty (تعليق a المسامير) levé sousedy o četnosti 10425 a 20. To je průměrně 5222,5. Pro pravé okolí je průměrná četnost 544.

Měření z kapitoly 5.2 se týká délky slova. Například levé okolí slov o délce 4 písmena (jsou to slova أرغب a وطأة) má délku 4 a 2 písmena. To jsou průměrně 3 písmena. Nebo například pravé okolí slov o 3 písmenech má délku 3,25.

Analogicky jsou měřeny vztahy, které se týkají vět.